TTI Success Insights
Emotional Quotient 2026 Technical Manual Version 2.1

August 5, 2026



TTI SUCCESS INSIGHTS®

List of Major Changes in the Draft

1 Table 1.1 . . . 2
2 Table 1.2 . . . 3
3 Section 1.3 . . . 4
4 Section 5 added . . . 56
5 Section 6 updated with new data . . . 66
6 Section 8 . . . 77
7 Section 9.6 added, Graded Response Models Summary . . . 104
8 Section 10 added Factor Analysis . . . 105
9 Section 10.8 added . . . 121
10 Section 11 updated with new data . . . 124
11 Appendix A . . . 168
12 Section B . . . 177
13 Section C.6 . . . 198
14 Section C.7 . . . 198
15 Appendix D . . . 199
16 Appendix E . . . 206
17 Appendix F . . . 213
18 Appendix G . . . 220
19 Appendix H . . . 227
20 Section I.6 . . . 240
21 Section I.7 . . . 240
22 Appendix J . . . 241
23 Appendix K . . . 248
24 Appendix L . . . 255
25 Appendix M . . . 262
26 Appendix N . . . 269
27 Appendix O . . . 277
28 Section P.6 . . . 290
29 Section P.7 . . . 290
30 Appendix Q . . . 291
31 Appendix R . . . 299

TTI SUCCESS INSIGHTS®

Table of Contents

List of Figures ix

List of Tables xi

List of Key Symbols xx

1 Introduction 1
- 1.1 Executive Summary . . . 1
- 1.2 How to Navigate this Manual . . . 2
- 1.3 Interpretation of the Content of This Manual . . . 4

2 A Short Summary of the History and Theory of Emotional Intelligence 6
- 2.1 Emotional Intelligence Theory Formalized . . . 6
- 2.2 The Bar-On Model of Emotional Intelligence . . . 8
- 2.3 The Emotional Competencies Model of Goleman . . . 9
- 2.4 A Comparison of the Salovey, Bar-On, and Goleman EI Models . . . 13
- 2.5 The Next Generation of EI Assessments . . . 14
- 2.6 Summary . . . 14

3 Our Approach to the Emotional Quotient Assessment 16
- 3.1 Goleman's 1995 EQ Model . . . 16
- 3.2 Goleman's 2000 EQ Model . . . 18
- 3.3 TTI Success Insights' Current EQ Model . . . 20
- 3.4 TTI Success Insights Ideal EQ Model . . . 21

4 A Few Comments on Reliability and Validity 39
- 4.1 Thoughts on Evidence of Reliability . . . 40
- 4.2 Thoughts on Evidence of Validity . . . 42
- 4.3 Correlation v. Causation . . . 48
- 4.4 Understanding Self-report Assessment Bias . . . 49
- 4.5 Forced-rank v. Likert-style Assessment Formats . . . 51
- 4.6 Parting Thoughts . . . 54

5 APA Guidance on Evidence of Reliability and Validity 56
- 5.1 Evidence of Reliability . . . 56
- 5.2 Evidence of Validity . . . 58
- 5.3 Deductive Versus Inductive Reasoning . . . 62
- 5.4 Summary . . . 63

6 Evidence of Reliability Based on Internal Consistency 66
- 6.1 A Short Review of the American Psychological Association's View of Reliability . . . 67
- 6.2 A Brief Discussion of Coefficient α . . . 67
- 6.3 α-if-item-deleted as a Measure of Internal Consistency . . . 67
- 6.4 Inter-item Correlation as a Measure of Internal Consistency . . . 68
- 6.5 Internal Consistency Results . . . 69
- 6.6 Summary . . . 72

7 Evidence of Reliability Based on Temporal Consistency 73
7.1 A Brief Review of the APA's View of Temporal Consistency . . . 73
7.2 Test-Retest Studies . . . 74
7.3 Test-Retest Summary . . . 75
7.4 Historical Test-Retest Results . . . 76

8 Evidence of Validity Based on Test Content 77
8.1 Construct Definition . . . 77
8.2 Construct Relevance . . . 79
8.3 Construct Representation . . . 80
8.4 Test Construction Procedures . . . 81
8.5 Historical Development . . . 83

9 Item Response Theory 86
9.1 From Classical Test Theory to Item Response Theory . . . 86
9.2 A Short Overview of the Rasch Models . . . 89
9.3 The Transition from Dichotomous to Free Response Models . . . 94
9.4 Item and Test Information . . . 98
9.5 Parameter Estimation and Scoring Rubric . . . 103
9.6 A Summary of the Graded Response Model Analysis . . . 104

10 Factor Analysis 105
10.1 Some Preliminaries . . . 105
10.2 A Brief History of Factor Analysis . . . 106
10.3 Geometric Interpretation of Correlation . . . 107
10.4 The Basic Factor Model . . . 113
10.5 Factor Analysis in Matrix Form . . . 116
10.6 A Condensed View of the MinRes Problem . . . 116
10.7 Factor Scores and Indeterminacy . . . 120
10.8 Weighted Scoring and Score Normalization . . . 121
10.9 Exploratory Factor Analytic Summary . . . 123

11 Evidence of Validity Based on Internal Structure 124
11.1 Corrected Item-to-Total Correlation . . . 124
11.2 Item Response Theory . . . 125
11.3 Exploratory Factor Analysis . . . 126
11.4 Internal Structure Current Results . . . 126
11.5 Summary . . . 134

12 Evidence of Validity Based on Relationships to External Variables 135
12.1 A Brief Review of Logistic Regression . . . 135
12.2 Confusion Matrices, ROC, and AUC . . . 137
12.3 A Brief Introduction to Principal Component Analysis . . . 141
12.4 Logistic Regression Study Results . . . 142

13 Response Processing Evidence of Validity 144
13.1 Using Electroencephalograph (EEG) to Measure ACI Response Processing . . . 144
13.2 Response Processing Summary . . . 147

14 Evidence of Validity Based on Consequences of Testing **148**
14.1 Additional Negative Consequences and Mitigation Efforts Acknowledged by TTI Success Insights . . . 148

15 Summary and Future Work **151**
15.1 Summary of Introductory Sections . . . 151
15.2 Summary of Evidence of Reliability Sections . . . 151
15.3 Summary of Evidence of Validity Sections . . . 152
15.4 Summary of our Appendices . . . 154
15.5 Future Work . . . 154

References **156**

A Historical Reliability Analysis **168**
A.1 General Population Results . . . 168
A.2 U.S. Historical Inter-Item Correlation Analysis . . . 170
A.3 U.S. Historical Internal Consistency via the α Coefficient . . . 173
A.4 U.S. Historical Internal Consistency via the α-if-item-deleted . . . 175

B Historical Internal Structure Analysis **177**
B.1 Historical Results . . . 177

C EQ Analysis for Australia **192**
C.1 Australian Inter-Item Correlation Analysis . . . 192
C.2 Australian Corrected Item Total Correlation Analysis . . . 193
C.3 Australian Internal Consistency via the α Coefficient . . . 194
C.4 Australian Test-Retest Analysis . . . 195
C.5 Australian Factor Analysis . . . 195
C.6 Australian EQ Model Equating Functions . . . 198
C.7 Australian EQ Model Summary . . . 198

D EQ Analysis for Brazil **199**
D.1 Brazilian Inter-Item Correlation Analysis . . . 199
D.2 Brazilian Corrected Item Total Correlation Analysis . . . 200
D.3 Brazilian Internal Consistency via the α Coefficient . . . 201
D.4 Brazilian Test-Retest Analysis . . . 202
D.5 Brazilian Factor Analysis . . . 202
D.6 Brazilian EQ Model Equating Functions . . . 205
D.7 Brazilian EQ Model Summary . . . 205

E EQ Analysis for Chinese **206**
E.1 Chinese Inter-Item Correlation Analysis . . . 206
E.2 Chinese Corrected Item Total Correlation Analysis . . . 207
E.3 Chinese Internal Consistency via the α Coefficient . . . 208
E.4 Chinese Test-Retest Analysis . . . 209
E.5 Chinese Factor Analysis . . . 209
E.6 Chinese EQ Model Equating Functions . . . 212
E.7 Chinese EQ Model Summary . . . 212

TTI SUCCESS INSIGHTS®

F EQ Analysis for Dutch **213**
F.1 Inter-Item Correlation Analysis . . . 213
F.2 Corrected Item Total Correlation Analysis . . . 214
F.3 Internal Consistency via the α Coefficient . . . 215
F.4 Test-Retest Analysis . . . 216
F.5 Factor Analysis . . . 216
F.6 Dutch EQ Model Equating Functions . . . 219
F.7 Dutch EQ Model Summary . . . 219

G EQ Analysis for French **220**
G.1 French Inter-Item Correlation Analysis . . . 220
G.2 French Corrected Item Total Correlation Analysis . . . 221
G.3 French Internal Consistency via the α Coefficient . . . 222
G.4 French Test-Retest Analysis . . . 223
G.5 French Factor Analysis . . . 223
G.6 French EQ Model Equating Functions . . . 226
G.7 French EQ Model Summary . . . 226

H EQ Analysis for French Canadien **227**
H.1 Inter-Item Correlation Analysis . . . 227
H.2 Item Total Correlation Analysis . . . 228
H.3 Internal Consistency via the α Coefficient . . . 229
H.4 Test-Retest Analysis . . . 230
H.5 Factor Analysis . . . 230
H.6 French Canadien EQ Model Equating Functions . . . 232
H.7 French Canadien EQ Model Summary . . . 232

I EQ Analysis for German **234**
I.1 Inter-Item Correlation Analysis . . . 234
I.2 Item Total Correlation Analysis . . . 235
I.3 Internal Consistency via the α Coefficient . . . 236
I.4 Test-Retest Analysis . . . 237
I.5 Factor Analysis . . . 237
I.6 German EQ Model Equating Functions . . . 240
I.7 German EQ Model Summary . . . 240

J EQ Analysis for Hungarian **241**
J.1 Inter Item Correlation Analysis . . . 241
J.2 Item Total Correlation Analysis . . . 242
J.3 Internal Consistency via the α Coefficient . . . 243
J.4 Test-Retest Analysis . . . 244
J.5 Factor Analysis . . . 244
J.6 Hungary EQ Model Equating Functions . . . 247
J.7 Hungarian EQ Model Summary . . . 247

K EQ Analysis for Italian **248**
K.1 Inter-Item Correlation Analysis . . . 248
K.2 Item-Total Correlation Analysis . . . 249

TTI SUCCESS INSIGHTS®

K.3 Internal Consistency via the α Coefficient . . . 250
K.4 Test-Retest Analysis . . . 251
K.5 Factor Analysis . . . 251
K.6 Italy EQ Model Equating Functions . . . 254
K.7 Italian EQ Model Summary . . . 254

L EQ Analysis for Portuguese 255
L.1 Portuguese Inter-Item Correlation Analysis . . . 255
L.2 Portuguese Corrected Item Total Correlation Analysis . . . 256
L.3 Portuguese Internal Consistency via the α Coefficient . . . 257
L.4 Portuguese Test-Retest Analysis . . . 258
L.5 Portuguese Factor Analysis . . . 258
L.6 Portuguese EQ Model Equating Functions . . . 261
L.7 Portuguese EQ Model Summary . . . 261

M EQ Analysis for Spain 262
M.1 Spanish ES Inter-Item Correlation Analysis . . . 262
M.2 Spanish ES Item Total Correlation Analysis . . . 263
M.3 Spanish ES Internal Consistency via the α Coefficient . . . 264
M.4 Spanish ES Test-Retest Analysis . . . 265
M.5 Spanish ES Factor Analysis . . . 265
M.6 Spanish ES EQ Model Equating Functions . . . 268
M.7 Spanish ES EQ Model Summary . . . 268

N EQ Analysis for Spanish North America 269
N.1 Spanish NA Inter-Item Correlation Analysis . . . 269
N.2 Spanish NA Corrected Item Total Correlation Analysis . . . 269
N.3 Spanish NA Internal Consistency via the α Coefficient . . . 272
N.4 Spanish NA Test-Retest Analysis . . . 273
N.5 Spanish NA Factor Analysis . . . 273
N.6 Spanish NA EQ Model Equating Functions . . . 276
N.7 Spanish NA EQ Model Summary . . . 276

O EQ Analysis for Swedish 277
O.1 Swedish Inter-Item Correlation Analysis . . . 277
O.2 Swedish Corrected Item Total Correlation Analysis . . . 278
O.3 Swedish Internal Consistency via the α Coefficient . . . 279
O.4 Swedish Test-Retest Analysis . . . 280
O.5 Swedish Factor Analysis . . . 280
O.6 Swedish EQ Model Equating Functions . . . 283
O.7 Swedish EQ Model Summary . . . 283

P EQ Analysis for the United Kingdom 284
P.1 UK Inter Item Correlation Analysis . . . 284
P.2 UK Corrected Item Total Correlation Analysis . . . 285
P.3 UK Internal Consistency via the α Coefficient . . . 286
P.4 UK Test-Retest Analysis . . . 287
P.5 UK Factor Analysis . . . 287

TTI SUCCESS INSIGHTS®

P.6 UK EQ Model Equating Functions 290
P.7 UK EQ Model Summary 290

Q Assessment Adaptation Protocols **291**
Q.1 Pre-conditions 292
Q.2 Test Development 293
Q.3 Confirmation Guidelines 294
Q.4 Administration Guidelines 296
Q.5 Score Scales and Interpretation Guidelines 297
Q.6 Documentation Guidelines 298

R Assessment Development Outline **299**

List of Figures

2.1 4-Branch Model of EI, Version I. . . . 7
2.2 4-Branch Model of EI, Version II 8
2.3 Bar-On's Model of EI 9
2.4 Goleman's 1995 EI Model 11
2.5 Comparison of 3 Initial EI Models 13
3.1 TTI Success Insights Current Emotional Intelligence Model 20
3.2 TTI Success Insights Emotional Intelligence Model 21
3.3 TTI Success Insights Self Awareness Construct 22
3.4 TTI Success Insights Emotional Self Awareness Sub-construct 23
3.5 TTI Success Insights Accurate Self Assessment Sub-construct 23
3.6 TTI Success Insights Self Confidence Sub-construct 24
3.7 TTI Success Insights Self Regulation Construct 25
3.8 TTI Success Insights Emotional Self Control Sub-construct 26
3.9 TTI Success Insights Trustworthiness Sub-construct 26
3.10 TTI Success Insights Conscientiousness Sub-construct 27
3.11 TTI Success Insights Adaptability Sub-construct 28
3.12 TTI Success Insights Achievement Drive Sub-construct 28
3.13 TTI Success Insights Initiative Drive Sub-construct 29
3.14 TTI Success Insights Social Awareness Construct 30
3.15 TTI Success Insights Empathy Sub-construct 30
3.16 TTI Success Insights Service Orientation Sub-construct 31
3.17 TTI Success Insights Service Orientation Sub-construct 32
3.18 TTI Success Insights Social Regulation Construct 33
3.19 TTI Success Insights Developing Others Sub-construct 33
3.20 TTI Success Insights Influence Sub-construct 34
3.21 TTI Success Insights Communication Sub-construct 35
3.22 TTI Success Insights Conflict Resolution Sub-construct 35
3.23 TTI Success Insights Visionary Leadership Sub-construct 36
3.24 TTI Success Insights Catalyzing Change Sub-construct 37
3.25 TTI Success Insights Building Bonds Sub-construct 37
3.26 TTI Success Insights Collaboration and Teamwork Sub-construct 38
4.1 Bad Correlation v Causation Example 48
9.1 Generic CDF 89
9.2 Example of Item Difficulty 90
9.3 Example of Item Discrimination 91
9.4 Rasch 1-PL Model 92
9.5 Rasch 2-PL Model 93
9.6 Rasch 3-PL Model 94
9.7 Example Boundary Characteristic Curves 99
9.8 Example Item Response Category Characteristic Curves 100
9.9 Example Item Information Function 101
9.10 Example Test Information Function 102
9.11 Example Item Information Function with Reliability Boundary 103
9.12 Example Test Information Function with Reliability Boundary 103
10.1 Vectors in $\mathbb{R}^n$ 108
10.2 Geometric Interpretation of Factor Analysis Parameters 120

TTI SUCCESS INSIGHTS®

11.1 Self Awareness GRM TIF 2026 . . . 128
11.2 Self Regulation GRM TIF 2026 . . . 128
11.3 Motivation GRM TIF 2026 . . . 129
11.4 Social Awareness GRM TIF 2026 . . . 129
11.5 Social Regulation GRM TIF 2026 . . . 130
12.1 Confusion Matrix . . . 138
12.2 Example ROC Curve . . . 140
13.1 An Example of Asymmetry in EEG Captured Brain Activity . . . 145
13.2 Confirmation Between the Survey and Brain Response ACI World View . . . 146
13.3 Confirmation Between the Survey and Brain Response ACI Self View . . . 147
B.1 Self Awareness GRM TIF 2025 . . . 179
B.2 Self Awareness GRM TIF 2025a . . . 179
B.3 Self Regulation GRM TIF 2025 . . . 180
B.4 Self Regulation GRM TIF 2025a . . . 180
B.5 Motivation GRM TIF 2025 . . . 181
B.6 Motivation GRM TIF 2025a . . . 181
B.7 Social Awareness GRM TIF 2025 . . . 182
B.8 Social Awareness GRM TIF 2025a . . . 182
B.9 Social Regulation GRM TIF 2025 . . . 183
B.10 Social Regulation GRM TIF 2025a . . . 183
Q.1 Emotional response to *enthusiastic* . . . 291

List of Tables

1.1 TTI Success Insights EQ Data . . . 2
1.2 TTI Success Insights EQ Technical Manual Navigation . . . 3
1.3 TTI Success Insights EQ Scaled Abbreviation Table . . . 4
2.1 Goleman's 2000 EI Model . . . 12
4.1 Common Reliability Coefficient Interpretation . . . 41
6.1 Common Reliability Coefficient Interpretation, see [63] . . . 69
6.2 U.S. General Population Statistics Standard Scoring N = 4,395 . . . 70
6.3 Average U.S. Inter-item Correlations . . . 71
6.4 Average U.S. α Coefficient Data . . . 71
6.5 α-if-item-deleted for U.S. . . . 72
7.1 2024 Test-Retest Data Breakdown TTI Success Insights EQ . . . 75
7.2 2024 Test-Retest Correlations TTI Success Insights EQ . . . 75
7.3 2018 Test-Retest Data Breakdown TTI Success Insights EQ . . . 76
7.4 Test-Retest Correlations 1/1/2011 to 12/14/2014 TTI Success Insights EQ . . . 76
7.5 Test-Retest Correlations 12/16/2014 to 5/7/2018 TTI Success Insights EQ . . . 76
9.1 Response Frequencies Dichotomization Example . . . 87
9.2 Response Distribution Dichotomization Example . . . 88
10.1 Examples of Observed Variables from [103] . . . 113
10.2 Composition of Variance for Full Common Factor Model . . . 115
11.1 Corrected Item-total Correlation Levels Suggested by APA . . . 125
11.2 US Corrected Item-total Correlations 2026 . . . 127
11.3 US Factor Pattern Matrix Self Awareness 2026 . . . 131
11.4 US Factor Pattern Matrix Self Regulation 2026 . . . 132
11.5 US Factor Pattern Matrix Motivation 2026 . . . 132
11.6 US Factor Pattern Matrix Social Awareness 2026 . . . 133
11.7 US Factor Pattern Matrix Social Regulation 2026 Current . . . 133
12.1 Example Confusion Matrix . . . 139
12.2 EQ Assemblers and Fabricators All Other Logit Regression Model, N = 113 . . . 142



12.3 Aortic Valve Sales Sample Logit Regression Model, N = 72 . . . 143
A.1 U.S. General Population Statistics Standard Scoring N = 4,395 . . . 168
A.2 U.S. General Population Statistics Weighted Scoring N = 4,395 . . . 168
A.3 U.S. General Population Statistics N = 3,803 . . . 168
A.4 U.S. General Population Statistics 2023 N = 87,704 . . . 169
A.5 Average U.S. Inter-item Correlations Weighted 2025 . . . 170
A.6 Average U.S. Inter-item Correlations 2025 Standard . . . 170
A.7 Average U.S. Inter-item Correlations 2023 . . . 171
A.8 Average U.S. Inter-item Correlations 2020 . . . 171
A.9 Average U.S. Inter-item Correlations 2018 . . . 172
A.10 Average U.S. α Coefficient Data 2025 Weighted . . . 173
A.11 Average U.S. α Coefficient Data 2025 Standard . . . 173
A.12 Average U.S. α Coefficient Data 2023 . . . 173
A.13 Average U.S. α Coefficient Data 2020 . . . 174
A.14 Average U.S. α Coefficient Data 2018 . . . 174
A.15 α-if-item-deleted for U.S. Weighted 2025 . . . 175
A.16 α-if-item-deleted for U.S. 2025 . . . 175
A.17 α-if-item-deleted for U.S. 2023 . . . 176
B.1 US Corrected Item-total Correlations 2025 Current . . . 177
B.2 US Corrected Item-total Correlations 2025 . . . 178
B.3 US Corrected Item-total Correlations 2023 . . . 178
B.4 US Factor Pattern Matrix Self Awareness 2025 . . . 184
B.5 US Factor Pattern Matrix Self Awareness 2025 . . . 184
B.6 US Factor Pattern Matrix Self Awareness 2023 . . . 185
B.7 US Factor Pattern Matrix Self Regulation 2025 . . . 185
B.8 US Factor Pattern Matrix Self Regulation 2025 . . . 186
B.9 US Factor Pattern Matrix Self Regulation 2023 . . . 186
B.10 US Factor Pattern Matrix Motivation 2025 . . . 187
B.11 US Factor Pattern Matrix Motivation 2025 . . . 187
B.12 US Factor Pattern Matrix Motivation 2023 . . . 188
B.13 US Factor Pattern Matrix Social Awareness 2025 . . . 188
B.14 US Factor Pattern Matrix Social Awareness 2025 . . . 189



B.15 US Factor Pattern Matrix Social Awareness 2023 . . . 190
B.16 US Factor Pattern Matrix Social Regulation 2025 . . . 190
B.17 US Factor Pattern Matrix Social Regulation 2025 . . . 191
B.18 US Factor Pattern Matrix Social Regulation 2023 . . . 191
C.1 Australian General Population Statistics N = 3380 . . . 192
C.2 Average Australian Inter-item Correlations . . . 192
C.3 Australian Corrected Item-total Correlations . . . 193
C.4 Common Reliability Coefficient Interpretation, see [63] . . . 194
C.5 Average Australian α Coefficient Data . . . 194
C.6 α-if-item-deleted for Australia . . . 195
C.7 Australian Factor Pattern Matrix Self Awareness . . . 195
C.8 Australian Factor Pattern Matrix Self Regulation . . . 196
C.9 Australian Factor Pattern Matrix Motivation . . . 196
C.10 Australian Factor Pattern Matrix Social Awareness . . . 197
C.11 Australian Factor Pattern Matrix Social Regulation . . . 197
C.12 Australian-U.S. Equating Functions Mean/StDev Method . . . 198
C.13 Australian-U.S. Equating Functions Equipercentile Method . . . 198
D.1 Brazilian General Population Statistics N = 4530 . . . 199
D.2 Average Brazilian Inter-item Correlations . . . 199
D.3 Brazilian Corrected Item-total Correlations . . . 200
D.4 Common Reliability Coefficient Interpretation, see [63] . . . 201
D.5 Average Brazilian α Coefficient Data . . . 201
D.6 α-if-item-deleted for Australia . . . 202
D.7 Brazilian Factor Pattern Matrix Self Awareness . . . 202
D.8 Brazilian Factor Pattern Matrix Self Regulation . . . 203
D.9 Brazilian Factor Pattern Matrix Motivation . . . 203
D.10 Brazilian Factor Pattern Matrix Social Awareness . . . 204
D.11 Brazilian Factor Pattern Matrix Social Regulation . . . 204

TTI SUCCESS INSIGHTS®

D.12 Brazilian-U.S. Equating Functions Mean/StDev Method . . . 205
D.13 Brazilian-U.S. Equating Functions N = 3380 . . . 205
E.1 Chinese General Population Statistics N = 1434 . . . 206
E.2 Average Chinese Inter-item Correlations . . . 206
E.3 Chinese Corrected Item-total Correlations TTI SI . . . 207
E.4 Common Reliability Coefficient Interpretation, see [63] . . . 208
E.5 Average Chinese α Coefficient Data TTI SI . . . 208
E.6 α-if-item-deleted for Chinese . . . 209
E.7 Chinese Factor Pattern Matrix Self Awareness . . . 209
E.8 Chinese Factor Pattern Matrix Self Regulation . . . 210
E.9 Chinese Factor Pattern Matrix Motivation . . . 210
E.10 Chinese Factor Pattern Matrix Social Awareness . . . 211
E.11 Chinese Factor Pattern Matrix Social Regulation . . . 211
E.12 Chinese-U.S. Equating Functions Mean/StDev Method . . . 212
E.13 Chinese-U.S. Equating Functions Equipercentile Method . . . 212
F.1 Dutch General Population Statistics N = 1647 . . . 213
F.2 Average Dutch Inter-item Correlations . . . 213
F.3 Dutch Corrected Item-total Correlations . . . 214
F.4 Common Reliability Coefficient Interpretation, see [63] . . . 215
F.5 Average Dutch α Coefficient Data . . . 215
F.6 α-if-item-deleted for Dutch . . . 216
F.7 Dutch Factor Pattern Matrix Self Awareness . . . 216
F.8 Dutch Factor Pattern Matrix Self Regulation . . . 217
F.9 Dutch Factor Pattern Matrix Motivation . . . 217
F.10 Dutch Factor Pattern Matrix Social Awareness . . . 218
F.11 Dutch Factor Pattern Matrix Social Regulation . . . 218
F.12 Dutch-U.S. Equating Functions Mean/StDev Method . . . 219

TTI SUCCESS INSIGHTS®

F.13 Dutch-U.S. Equating Functions Equipercentile Method . . . 219
G.1 French General Population Statistics N = 4146 . . . 220
G.2 Average French Inter-item Correlations . . . 220
G.3 French Corrected Item-total Correlations . . . 221
G.4 Common Reliability Coefficient Interpretation, see [63] . . . 222
G.5 Average Frencha α Coefficient Data . . . 222
G.6 α-if-item-deleted for France . . . 223
G.7 French Factor Pattern Matrix Self Awareness . . . 223
G.8 French Factor Pattern Matrix Self Regulation . . . 224
G.9 French Factor Pattern Matrix Motivation . . . 224
G.10 French Factor Pattern Matrix Social Awareness . . . 225
G.11 French Factor Pattern Matrix Social Regulation . . . 225
G.12 French-U.S. Equating Functions Mean/StDev Method . . . 226
G.13 French-U.S. Equating Functions Equipercentile Method . . . 226
H.1 French Canadien General Population Statistics N = 162 . . . 227
H.2 Average French Canadien Inter-item Correlations . . . 227
H.3 French Canadien Corrected Item-total Correlations . . . 228
H.4 Common Reliability Coefficient Interpretation, see [63] . . . 229
H.5 Average French Canadien α Coefficient Data . . . 229
H.6 α-if-item-deleted for French Canadien . . . 230
H.7 French Canadien Factor Pattern Matrix Self Awareness . . . 230
H.8 French Canadien Factor Pattern Matrix Self Regulation . . . 231
H.9 French Canadien Factor Pattern Matrix Motivation . . . 232
H.10 French Canadien Factor Pattern Matrix Social Awareness . . . 232
H.11 French Candien Factor Pattern Matrix Social Regulation . . . 233
H.12 French Canadien-U.S. Equating Functions Mean/StDev Method . . . 233
H.13 French Canadien-U.S. Equating Functions Equipercentile Method . . . 233
I.1 German General Population Statistics N = 8004 . . . 234

TTI SUCCESS INSIGHTS®

I.2 Average German Inter-item Correlations . . . 234
I.3 German Corrected Item-total Correlations . . . 235
I.4 Common Reliability Coefficient Interpretation, see [63] . . . 236
I.5 Average German α Coefficient Data . . . 236
I.6 α-if-item-deleted for German . . . 237
I.7 German Factor Pattern Matrix Self Awareness . . . 237
I.8 German Factor Pattern Matrix Self Regulation . . . 238
I.9 German Factor Pattern Matrix Motivation . . . 238
I.10 German Factor Pattern Matrix Social Awareness . . . 239
I.11 German Factor Pattern Matrix Social Regulation . . . 239
I.12 German-U.S. Equating Functions Mean/StDev Method . . . 240
I.13 German-U.S. Equating Functions Equipercentile Method . . . 240
J.1 Hungarian General Population Statistics N = 342 . . . 241
J.2 Average Hungarian Inter-item Correlations . . . 241
J.3 Hungarian Corrected Item-total Correlations . . . 242
J.4 Common Reliability Coefficient Interpretation, see [63] . . . 243
J.5 Average Hungarian α Coefficient Data . . . 243
J.6 α-if-item-deleted for Hungarian . . . 244
J.7 Hungarian Factor Pattern Matrix Self Awareness . . . 244
J.8 Hungarian Factor Pattern Matrix Self Regulation . . . 245
J.9 Hungarian Factor Pattern Matrix Motivation . . . 245
J.10 Hungarian Factor Pattern Matrix Social Awareness . . . 246
J.11 Hungarian Factor Pattern Matrix Social Regulation . . . 246
J.12 Hungary-U.S. Equating Functions Mean/StDev Method . . . 247
J.13 Hungary-U.S. Equating Functions Equipercentile Method . . . 247
K.1 Italian General Population Statistics N = 848 . . . 248
K.2 Average Italian Inter-item Correlations . . . 248
K.3 Italian Corrected Item-total Correlations . . . 249
K.4 Common Reliability Coefficient Interpretation, see [63] . . . 250



K.5 Average Italian α Coefficient Data . . . 250
K.6 α-if-item-deleted for Italian . . . 251
K.7 Italian Factor Pattern Matrix Self Awareness . . . 251
K.8 Italian Factor Pattern Matrix Self Regulation . . . 252
K.9 Italian Factor Pattern Matrix Motivation . . . 252
K.10 Italian Factor Pattern Matrix Social Awareness . . . 253
K.11 Italian Factor Pattern Matrix Social Regulation . . . 253
K.12 Italy-U.S. Equating Functions Mean/StDev Method . . . 254
K.13 Italy-U.S. Equating Functions Equipercentile Method . . . 254
L.1 Portuguese General Population Statistics N = 180 . . . 255
L.2 Average Portuguese Inter-item Correlations . . . 255
L.3 Portuguese Corrected Item-total Correlations . . . 256
L.4 Common Reliability Coefficient Interpretation, see [63] . . . 257
L.5 Average Portuguese α Coefficient Data . . . 257
L.6 α-if-item-deleted for Portuguese . . . 258
L.7 Portuguese Factor Pattern Matrix Self Awareness . . . 258
L.8 Portuguese Factor Pattern Matrix Self Regulation . . . 259
L.9 Portuguese Factor Pattern Matrix Motivation . . . 259
L.10 Portuguese Factor Pattern Matrix Social Awareness . . . 260
L.11 Portuguese Factor Pattern Matrix Social Regulation . . . 260
L.12 Portuguese-U.S. Equating Functions Mean/StDev Method . . . 261
L.13 Portuguese-U.S. Equating Functions Equipercentile Method . . . 261
M.1 Spanish ES General Population Statistics N = 2284 . . . 262
M.2 Average Spanish ES Inter-item Correlations TTI SI Emotional Quotient . . . 262
M.3 Spanish ES Corrected Item-total Correlations . . . 263
M.4 Common Reliability Coefficient Interpretation, see [63] . . . 264
M.5 Average Spanish ES α Coefficient Data TTI SI Emotional Quotient . . . 264
M.6 α-if-item-deleted for Spanish ES . . . 265



M.7 Spanish ES Factor Pattern Matrix Self Awareness . . . 265
M.8 Spanish ES Factor Pattern Matrix Self Regulation . . . 266
M.9 Spanish ES Factor Pattern Matrix Motivation . . . 266
M.10 Spanish ES Factor Pattern Matrix Social Awareness . . . 267
M.11 Spanish ES Factor Pattern Matrix Social Regulation . . . 267
M.12 Spanish ES-U.S. Equating Functions Mean/StDev Method . . . 268
M.13 Spanish ES-U.S. Equating Functions Equipercentile Method . . . 268
N.1 Spanish NA General Population Statistics N = 2818 . . . 269
N.2 Average Spanish NA Inter-item Correlations . . . 270
N.3 Spanish Corrected Item-total Correlations . . . 270
N.4 Common Reliability Coefficient Interpretation, see [63] . . . 272
N.5 Average Spanish NA α Coefficient Data . . . 272
N.6 α-if-item-deleted for Spanish NA . . . 273
N.7 Spanish NA Factor Pattern Matrix Self Awareness . . . 273
N.8 Spanish NA Factor Pattern Matrix Self Regulation . . . 274
N.9 Spanish NA Factor Pattern Matrix Motivation . . . 274
N.10 Spanish NA Factor Pattern Matrix Social Awareness . . . 275
N.11 Spanish NA Factor Pattern Matrix Social Regulation . . . 275
N.12 Spanish NA-U.S. Equating Functions Mean/StDev Method . . . 276
N.13 Spanish NA-U.S. Equating Functions Equipercentile Method . . . 276
O.1 Swedish General Population Statistics N = 1092 . . . 277
O.2 Average Swedish Inter-item Correlations . . . 277
O.3 Swedish Corrected Item-total Correlations . . . 278
O.4 Common Reliability Coefficient Interpretation, see [63] . . . 279
O.5 Average Swedish α Coefficient Data . . . 279
O.6 α-if-item-deleted for Swedish . . . 280
O.7 Swedish Factor Pattern Matrix Self Awareness . . . 280
O.8 Swedish Factor Pattern Matrix Self Regulation . . . 281



O.9 Swedish Factor Pattern Matrix Motivation . . . 281
O.10 Swedish Factor Pattern Matrix Social Awareness . . . 282
O.11 Swedish Factor Pattern Matrix Social Regulation . . . 282
O.12 Swedish-U.S. Equating Functions Mean/StDev Method . . . 283
O.13 Swedish-U.S. Equating Functions Equipercentile Method . . . 283
P.1 UK General Population Statistics N = 1884 . . . 284
P.2 Average English-UK Inter-item Correlations . . . 284
P.3 English-UK Corrected Item-total Correlations . . . 285
P.4 Common Reliability Coefficient Interpretation, see [63] . . . 286
P.5 Average UK α Coefficient Data . . . 286
P.6 α-if-item-deleted for English UK . . . 287
P.7 UK Factor Pattern Matrix Self Awareness . . . 287
P.8 UK Factor Pattern Matrix Self Regulation . . . 288
P.9 UK Factor Pattern Matrix Motivation . . . 288
P.10 UK Factor Pattern Matrix Social Awareness . . . 289
P.11 UK Factor Pattern Matrix Social Regulation . . . 289
P.12 U.K.-U.S. Equating Functions Mean/StDev Method . . . 290
P.13 U.K.-U.S. Equating Functions Equipercentile Method . . . 290

TTI SUCCESS INSIGHTS®

List of Key Symbols

EI	Emotional Intelligence	6
IQ	Intelligence Quotient	6
α	Coefficient α	67
α_{se}	standardized estimate of α	69
MTT	Modern Test Theory	86
ρ_{XT}	classical version of test reliability	87
CDF	cumulative density function	89
$\mathbb{R}$	the real line	90
∞	infinity symbol	90
θ_n	ability level of the n^{th} respondent	92
δ_i	location of the i^{th} item	92
$Pr\{X_{ni} = 1\|\theta_n\}$	conditional probability of choosing the correct answer	92
$M_{ij}(\theta)$	processing function	96
$P^*_{i_j}(\theta)$	probability of responding positively at a boundary	96
P_{i_j}	difference in consecutive boundary probabilities	96
CBRF	category boundary response function	97
ICRF	item characteristic response function	97
IIF	item information function	98
TIF	test information function	98
BCC	boundary characteristic curve	99
IRCCC	item response category characteristic curves	99
I_{ij}	information from the i^{th} item associated with the j^{th} response category	100
$I_i(\theta)$	item information	100
λ	vector of factor loadings	105
$\mathbb{1}$	vector of all ones of appropriate length	105
$\mathbb{R}^n$	n-dimensional Euclidean space	108
$\cdot$	the usual dot product in $\mathbb{R}^n$	108
$\mu(X)$	expectation of the random variable X	109
$\bar{X}$	the mean or average of X	109
$\sigma^2(X)$	the variance of X	109
$\sigma(X, Y)$	the covariance between X and Y	111
$\rho_{X,Y}$	the correlation between X and Y	111
R^2	coefficient of determination	113
ITC	Item-Total Correlation	124

TTI SUCCESS INSIGHTS®

1 Introduction

The following manual contains information on the history and development of emotional intelligence from its beginnings to current implementation of the concepts as psychometric assessments. Included in this manual are a host of statistical and psychometric analyses used to establish evidence of the validity and reliability of the TTI Success Insights EQ assessment.

1.1 Executive Summary

The *TTI Success Insights Emotional Quotient Technical Manual* provides a comprehensive analysis of the development, validity, and reliability of the TTI Success Insights Emotional Quotient (EQ) assessment. This manual serves as both a theoretical and technical guide, detailing the evolution of emotional intelligence (EI) theory and the methodologies used to assess and refine the EQ model.

The document is structured to present the history and scientific foundation of emotional intelligence, followed by the psychometric processes used to validate the assessment. The TTI Success Insights EQ model is built upon five core constructs[1]:

- Self Awareness (SA) - recognizing and understanding one's emotions.
- Self Regulation (SR) - managing one's emotions effectively.
- Motivation (MO) - utilizing emotions to drive goal-oriented behavior.
- Social Awareness (SoA) - understanding and responding to the emotions of others.
- Social Regulation (SoR) - navigating social interactions and relationships effectively.

This manual outlines key psychometric principles, including:

- Item Analysis - evaluating the item difficulty and discrimination of the assessment items to ensure assessment accuracy.
- Reliability - various methods to measure test-retest reliability (correlations) and internal consistency (the so-called Cronbach's α coefficient, inter-item correlations) ensuring consistency of responses within the relevant population.
- Internal Structure Validity - various methods to measure the validity evidence based on internal structure, such as corrected item-to-total correlation, exploratory factor analysis, and graded response models from item response theory, ensuring accuracy of the model based on the relevant population.
- The Remaining Validity Categories - all five categories of evidence of validity, as defined by the American Psychological Association, are presented within this manual.

Additionally, the manual presents updates based on large-scale data analysis across multiple languages, demonstrating the continual improvement process applied to the EQ assessment. It also explores future research directions, including the integration of neuroscientific methods to enhance the understanding of emotional intelligence.

This technical manual is a valuable resource for researchers, practitioners, and organizations seeking an evidence-based approach to measuring and developing emotional intelligence in individuals and teams.

[1] These are not the construct definitions. Rather, these are presented as layperson descriptions of the concepts. Concrete definitions are provided in Sections 2 and 3

1.2 How to Navigate this Manual

This manual is presented in a manner consistent with development of a psychometric assessment from the theoretical foundation through final factor analysis to confirm the assessment performs as desired. Since the TTI Success Insights assessment family covers many assessments across multiple languages, the analyses are required to be performed on all combinations of languages and assessments. For this reason, this manual presents the main body of work in the language of development, in this case North American English.

In an ideal situation, the assessment in question would go through a process, possibly modified from the original development process, to generate an assessment based on the best set of items for a given language. In many cases, there does not exist enough data to support full factor analytic approaches to the final pieces of the development process. It is the goal of TTI Success Insights to support such assessment development wherever possible in future versions of our assessments. This approach provides a stronger case for evidence of validity of the assessment in all languages. Translation of an assessment from one language to another cannot guarantee transference of reliability and validity regardless of the starting point.

Table 1.1: TTI Success Insights EQ Data

Language	Total Data Analyzed
United States (North America)	87,704
Australia/New Zealand	3,380
Brazil	4,530
Chinese	1,434
Dutch	1,647
French	4,146
French Canadien	162
German	8,004
Hungarian	342
Italian	848
Portuguese	180
Spanish	2,284
Spanish North America	2,818
Swedish	1,092
United Kingdom	1,884

Table 1.1 presents the number of data points used in the analysis of the TTI Success Insights EQ assessment for each language[2] as well as showing the date that version of the assessment was most recently updated. The TTI Success Insights business model has historically been such that collection of most usual demographic information was viewed as invasive and the practice was not followed. Over time, this view point has migrated and the collection of basic demographic information is now deemed highly important to understanding the performance of our assessments, as well as laying the foundation for a highly effective continuous improvement process. Future

[2]Two separate data samples are used in the analysis of English in North America. The larger sample here represents historical data. The other is based on an updated version of the questionnaire and is first presented in Section 6.

versions of this technical manual are scheduled to include demographic information and analyses based on demographic breakdowns.

Table 1.2: TTI Success Insights EQ
Technical Manual Navigation

Language	IIC	ITC	α	T/R	EFA	IRT
US (North America)	6.5.1	11.4.1	6.5.2	7	11.4.3	11.4.2
AU/NZ	C.1	C.2	C.3	C.4	C.5	N/A
Brazil	D.1	D.2	D.3	D.4	D.5	N/A
Chinese	E.1	E.2	E.3	E.4	E.5	N/A
Dutch	F.1	F.2	F.3	F.4	F.5	N/A
French	G.1	G.2	G.3	G.4	G.5	N/A
French Canadien	H.1	H.2	H.3	H.4	H.5	N/A
German	I.1	I.2	I.3	I.4	I.5	N/A
Hungarian	J.1	J.2	J.3	J.4	J.5	N/A
Italian	K.1	K.2	K.3	K.4	K.5	N/A
Portuguese	L.1	L.2	L.3	L.4	L.5	N/A
Spanish	M.1	M.2	M.3	M.4	M.5	N/A
Spanish NA	N.1	N.2	N.3	N.4	N.5	N/A
Swedish	O.1	O.2	O.3	O.4	O.5	N/A
UK	P.1	P.2	P.3	P.4	P.5	N/A

Table 1.2 presents a shortcut to the main analyses presented in this manual, signified by the blue highlighted text. Throughout this manual, blue highlighted text represents hyperlinks internal to the document. The abbreviations used in the column headers for Table 1.2 are as follows.

IIC represents the average inter item correlation, see Section 6.4 for a detailed explanation. ITC represents the average (corrected) item-to-total-score correlation, see Section 11.4.1 for a detailed explanation. α (the Greek letter alpha) represents the α coefficient usually incorrectly attributed to Cronbach, see [58]. The α-coefficient and a closely related analysis known as α-if-item-deleted are discussed in Section 6.5.2. T/R refers to the section on test-retest or temporal consistency, a second part of the presentation on reliability in this manual.

We also present exploratory factor analysis (EFA) in Section 11.4.3. The final analytic tool discussed in this manual is item response theory (IRT). IRT is presented in Section 11.4.2. Note that only North American (US) English is analyzed using IRT at this time. The reason for this is simply that most languages outside North America fail to have large enough sample sizes to properly analyze at this time. Further, given that we are in the process of some significant improvements to the English version, the next step is to work through a translation and adaptation process for our global network. Subsequent to that occurring, we will outline and design a data capture and analytic approach that will incorporate IRT at the appropriate time.

The Table of Contents, List of Figures, List of Tables, and List of Key Symbols all contain hyperlinks to the relevant sections, figures, tables, or definitions of the key symbols. If the reader is using the Adobe Acrobat PDF Viewer, all blue highlighted text signifies a hyperlink to an appropriate portion of the document. This includes all bibliographic references presented throughout the text and detailed in the References section at the end of this document. In addition to all blue highlighted text signifying hyperlinks, hovering over the blue text generates a preview of the referenced material.

As a final note, the following abbreviations are used throughout the document to refer to the scales or constructs of the TTI Success Insights EQ Assessment.

Table 1.3: TTI Success Insights EQ
Scaled Abbreviation Table

Scale	Abbrev.
Self Awareness	SA
Self Regulation	SR
Motivation	MO
Social Awareness	SoA
Social Regulation	SoR

1.3 Interpretation of the Content of This Manual

This manual is laid out in four basic areas. These are the Introductory sections, the sections on evidence of reliability, the sections on evidence of validity, and the appendices. We present a brief synopsis of each for convenience. The reader interested only in the final discussion may consult the summary and future work section presented in Section 15.

We consider the introductory sections to be Sections 1 through 5. The first section is, obviously, the actual introduction presented here. After that, Section 2 presents a synopsis of the historical development of the EI model from multiple perspectives. Section 3 gives a presentation of the TTI Success Insights implementation of our EQ model. This is followed by our commentary on how we view the current state of the psychometric analysis, see Section 4. We finish the introductory sections with an outline of the American Psychological Association's (APA) guidance on establishing arguments for evidence of reliability and validity in Section 5.

We present two main sections covering our work on establishing our argument for evidence of reliability in Sections 6 and 7. The sections discuss internal consistency reliability and test-retest reliability, respectively. We discuss our reasons for not including work on alternate forms reliability and generalizability theory. The basic idea behind the exclusions are related to one being cost prohibitive and the other requiring a level of granularity in our data which just does not exist at this time.

We next dive into the areas related to establishing evidence of validity arguments. Up first is Section 8, where we discuss our approach to establishing a test content validity evidence argument. Before diving into internal structure validity evidence in Section 11, we present some theoretical background on IRT and EFA in sections 9 and 10, respectively. For those readers who are particularly interested in data analysis and the mathematics and statistics involved, the aforementioned three sections are a definite read. In particular, Section 11 ultimately presents the meat of the analytic results for validity arguments in this work. The other areas to establish validity arguments are largely logical progression arguments and processes.

Section 12 outlines one of our approaches for aiding in establishing evidence of validity based on relationships to external variables. The reader familiar with earlier versions of psychometric validity approaches may understand this area best as a combination of the previously known concurrent and predictive validity. We finish our discussions with Section 13 on response processing, where we present our neurological research based approach, and Section 14, where we, in part, apply



some of our neurological findings during our approach to consequences of testing. The manual is summarized in Section 15 which contains a synopsis and an outline of some thoughts on future directions.

We finish the manuscript with a series of appendices. Appendices C through P contain the results and following discussions of each language into which we have our EQ assessment translated and adapted. Appendix Q is a synopsis of the International Test Commission's guidance on translation and adaptation upon which we base our internal approach to adaptation. Finally, Appendix R contains a schematic of our overall assessment development and continual improvement process.

2 A Short Summary of the History and Theory of Emotional Intelligence

The purpose of this section is to lay a foundation for understanding the theory underlying the concept of emotional intelligence (EI) and to describe how this led to the development of the TTI Success Insights Emotional Quotient assessment. In order to lay such a foundation, it is first necessary to understand some of the history that lies at the heart of the theoretical development of emotional intelligence. It is well beyond the scope of this document to trace the entire history and development of this topic. TTI Success Insights has attempted to provide a reasonably sound outline of the history and theory while attaching a comprehensive list of references for the interested reader.

Before beginning down the path of the history and theoretical development, it may be helpful to first define what one means when using the term emotional intelligence. As an example, the authors in [153] offer the following definition.

Definition 2.1. *Emotional Intelligence is the ability to perceive emotions, to access and generate emotions so as to assist thought, to understand emotions and emotional meanings, and to reflectively regulate emotions so as to promote better emotion and thought.*

Many versions of a definition of EI have been put forth over the years. The remainder of this section outlines a handful of the more pervasive and respected versions. These versions trace out some of the history and development of the concepts beginning with Salovey and Mayer, through Bar-On, and onto Goleman. Each has a slightly different take on both the concept of emotional intelligence and how to measure it.

2.1 Emotional Intelligence Theory Formalized

As noted previously, the purpose of this section is to review the theoretical underpinnings of EI to develop an historical perspective that will lay the groundwork for how one might ultimately measure the aforementioned, albeit as yet unnamed, constructs.

In the 1980s, psychologists began to explore the role of emotions on thought and decision-making. A number of studies focused on how mood states influence mental processing and judgments, see [118, 134, 149]. The first references to the concept of emotions having a spectrum of ability, much like the early definitions of intelligence, can be traced to [128] which states

> emotions ... do not at all deserve being put into opposition with 'intelligence' ... they are, it seems, themselves a high order of intelligence (pp. 307-308).

Payne appears to be the first to have coined the term emotional intelligence, using it in his unpublished doctoral dissertation in 1983, see [137]. Bar-On, [13], added to the story by presenting the term emotional quotient (EQ) in his doctoral dissertation as a counterpart to intelligence quotient (IQ), or cognitive ability. Bar-On described EQ as a set of social and emotional abilities that help individuals cope with the demands of daily life. While all of these early works laid the groundwork for EI, none of them provided an organized model.

In 1990, the authors in [150] presented a framework and formal definition describing EI as

> ... a form of social intelligence that involves the ability to monitor one's own and others' feelings and emotions, to discriminate among them, and to use this information to guide one's own thinking and action (pp189).

Later during the same year, [120] followed with an empirical demonstration of how EI could be assessed as a unique mental ability. These foundational references describe three core components of EI: **Appraisal and Expression**, **Regulation**, and **Utilization**.

The emerging Mayer and Salovey model drew from not only their own research, but also a large body of theoretical research that was occurring during this time including: Averill & Nunley in [6], Buck in [37], Lane, et. al., in [106], Mayer & Geher in [121], Mayer & Stevens in [124], and Rosenthal, et. al. in [148]. In 1997, Salovey and Sluyter, see [153], introduced the four-dimensional model that forms the basis for many of today's EI assessments by conceptualizing EI as a set of abilities that are measurable.

The Salovey-Sluyter model ultimately introduced the core constructs that build on each other and included sub-developmental components as well. The first theme, called **Emotional Perception and Expression**, entails recognizing and inputting both verbal and nonverbal information from the emotional centers of the brain. The second theme, titled **Emotional Facilitation of Thought**, involves using emotions as part of the cognitive process. The third theme, called **Emotional Understanding**, involves cognitive processing that is insightful regarding one's own feelings or the feelings of others. The fourth and final Salovey-Sluyter construct is **Emotional Management** and outlines the steps required for regulation of self and others.

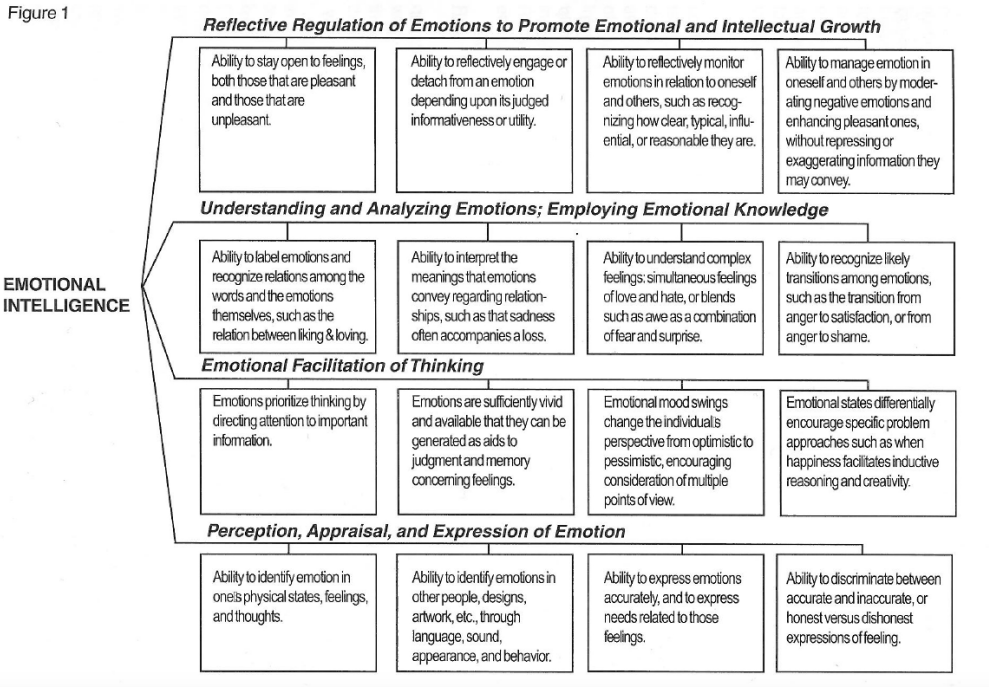


Figure 2.1: 4-Branch Model of EI, Version I.

The Mayer, Salovey, and Caruso model was not the only attempt at creating a framework for EI. While the Mayer, Salovey, and Caruso model focused on identifying stages of abilities within four domains and is referred to as an Ability Model, Bar-On, see [14] and [15], is classified as a Trait Model, see Section 2.2, and Goleman [79] describes a Competency Model, see Section 2.3. One may consult [135] for a more detailed discussion. Both the Bar-on and Goleman models are also referred to as mixed models.

The Mayer-Salovey model was initially known as the Four-Branch Model of Emotional Intelligence.

TTI SUCCESS INSIGHTS®

The four components, referred to as branches of EI, are further elaborated on in Figure 2.1, directly reproduced from [122]. To establish the groundwork for building an assessment around these four branches, Salovey, et. al., [151], published a description of the associated abilities tied to each branch, see Figure 2.2.

1. Emotional Perception and Expression
   (a) Ability to identify emotion in one's physical and psychological states
   (b) Ability to identify emotion in other people
   (c) Ability to express emotions accurately and to express needs related to them
   (d) Ability to discriminate between accurate/honest and inaccurate/dishonest feelings
2. Emotional Facilitation of Thought (Using Emotional Intelligence)
   (a) Ability to redirect and prioritize thinking on the basis of associated feelings
   (b) Ability to generate emotions to facilitate judgment and memory
   (c) Ability to capitalize on mood changes to appreciate multiple points of view
   (d) Ability to use emotional states to facilitate problem-solving and creativity
3. Emotional Understanding
   (a) Ability to understand relationships among various emotions
   (b) Ability to perceive the causes and consequences of emotions
   (c) Ability to understand complex feelings, emotional blends, and contradictory states
   (d) Ability to understand transition among emotions
4. Emotional Management
   (a) Ability to be open to feelings, both pleasant and unpleasant
   (b) Ability to monitor and reflect on emotions
   (c) Ability to engage, prolong, or detach from an emotional state
   (d) Ability to manage emotions in oneself
   (e) Ability to manage emotions in others

Figure 2.2: 4-Branch Model of EI, Version II

2.2 The Bar-On Model of Emotional Intelligence

Bar-on's model, see [14], defines EI as "an array of non-cognitive capabilities, competencies, and skills that influence one's ability to succeed in coping with environmental demands and pressures'.' By blending emotions and social competencies, this assessment is best described as an *emotional and social intelligence model*. As is shown in Figure 2.3, each of the five constructs is subdivided for a total of 15 measurable components.

Bar-On and Salovey, et. al., agree that EI may be improved through training and development programs. Hence, it is possible for one's EI to change throughout one's life. However, unlike the model of Salovey, et. al., Bar-On's model relates potential for performance rather than the performance itself. Bar-On's model was designed to address the question: *Why are some individuals more able to succeed in life than others?* By reviewing the psychological personality literature

TTI SUCCESS INSIGHTS®

related to success, he expanded on Salovey's original concept to include non-ability traits. Bar-On provided a rationale for his expanded definition of EI:

> Intelligence describes the aggregate of abilities, competencies, and skills . . . that . . . represent a collection of knowledge used to cope with life effectively. The adjective emotional is employed to emphasize that this specific type of intelligence differs from cognitive intelligence . . . [14], p. 15.

1. Intrapersonal Skills
   (a) Emotional Self Awareness
   (b) Assertiveness
   (c) Self Regard
   (d) Self Actualization
   (e) Independence
2. Interpersonal Skills
   (a) Interpersonal Relationships
   (b) Social Responsibility
   (c) Empathy
3. Adaptability
   (a) Problem Solving
   (b) Reality Testing
   (c) Flexibility
4. Stress Management
   (a) Stress Tolerance
   (b) Impulse Control
5. General Mood
   (a) Happiness
   (b) Optimism

Figure 2.3: Bar-On's Model of EI

The most well-known adaptation of the Bar-On Emotional Quotient Inventory is the EQ-i, an assessment that is constructed based on the Bar-On model and closely follows the primary psychometric constructs and sub-constructs outlined in Figure 2.3.

2.3 The Emotional Competencies Model of Goleman

Daniel Goleman was a science writer for the New York Times for 12 years where he specialized in brain and human behavior research. He trained as a psychologist at Harvard University where he worked with David McClelland, among others. During the 1990s, Goleman was introduced to the work of Salovey and Mayer, which ultimately lead to the publication of two books, "Emotional Intelligence", [79], and "Working with Emotional Intelligence", [81]. While Salovey and Mayer

laid the theoretical foundation for emotional intelligence, Goleman is credited for popularizing the concept with his marketing flare and prolific writing abilities, see, e.g., [42], [43], [67], [79], [80], [81], [82], [83], [84],[85], [86], [87], and [88].

Goleman's EI model was designed specifically for workplace applications and focuses on a wide array of managerial-related performance competencies and skills. [81] explores the function of EI on the job and claimed EI to be the strongest predictor of success in the workplace. This rather bold statement was based on findings from a world-wide sample summarized in [32]. In their study of nearly 200 large, global companies representing more than 500,000 individuals, Goleman states that truly effective leaders are distinguished by a high degree of emotional intelligence. He goes on to state that without well-developed EI, a person can have first-class training and an incisive mind, but still won't be a great leader.

Goleman states in [79] that, "at best, IQ contributes about 20 percent to the factors that determine life success, which leaves 80 percent to other factors." He further states that EI may be as powerful as, and at times more powerful than, IQ. Even Mayer, et. al., [123], state that "It is hard not to conclude that at least part of the popular excitement surrounding emotional intelligence is due to these very strong claims."

Over the years, even though Goleman does not, in fact, state that emotional intelligence represents the remaining 80 percent of life success, many filled in this gap and flocked to find ways to uncover a magic bullet. These optimistic views and their diffusion to the mass media opened a Pandora's box leading to promises of improved performance and professional success. However, empirical support for these statements over two decades later are still being sought, see, e.g., [3, 91, 100, 180].

This is not to say that EI is not a major factor in success. Rather, it is an indicator that a) there is no doubt that EI is one of many personal attributes that may account for the remaining 80 percent of life success and b) further research is necessary to determine how much variance in success may be attributed to EI. As recently as November, 2011, Goleman dealt with this misunderstanding by writing in his blog

> Have you heard? They say that your EQ counts for more than IQ for success. In fact, they say, EQ accounts for 80% of success. As the person who wrote Emotional Intelligence, the book that put this concept on the map, I can tell you they are dead wrong.

He then clarifies his stance by stating

> Here are the facts. There's no question IQ is by far the better determinant of career success, in the sense of predicting what kind of job you will be able to hold. It typically takes an IQ about 115 or above to be able to handle the cognitive complexity facing an accountant, physician, or a top executive. But here's the paradox: once you're in a high-IQ position, intellect loses its power to determine who will emerge as a productive employee or an effective leader. For that, how you handle yourself and your relationships - in other words, the emotional skill set - matters more than your IQ. In a high-IQ job pool, soft skills like discipline, drive, and empathy mark those who emerge as outstanding.

While researchers may question the specific contribution of EI to workplace success, an internet search at the time of this publication found over 19 million references to emotional intelligence and over 541,000 references to the corresponding term emotional quotient. Goleman took the work of

Mayer and Salovey and not only laid the groundwork for emotional intelligence to be a household concept, but also built a new theory-base.

1. **Self Awareness** - The ability to recognize and understand your moods, emotions, and drives, as well as their effect on others.
   (a) Emotional Awareness
   (b) Accurate Self Assessment
   (c) Self Confidence
2. **Self Regulation** - The ability to control or redirect disruptive impulses and moods. The propensity to suspend judgment - to think before acting.
   (a) Self Control
   (b) Trustworthiness
   (c) Conscientiousness
   (d) Adaptability
   (e) Innovation
3. **Motivation** - A passion to work for reasons that go beyond money or status. A propensity to pursue goals with energy and persistence.
   (a) Achievement Drive
   (b) Commitment
   (c) Initiative
   (d) Optimism
4. **Empathy** - The ability to understand the emotional makeup of other people. Skill in treating people according to their emotional reactions.
   (a) Understanding Others
   (b) Developing Others
   (c) Service Orientation
   (d) Leveraging Diversity
   (e) Political Awareness
5. **Social Skills** - Proficiency in managing relationships and building networks. An ability to find common ground and build rapport.
   (a) Influence
   (b) Communication
   (c) Conflict Management
   (d) Leadership
   (e) Change Catalyst
   (f) Building Bonds
   (g) Collaboration and Cooperation
   (h) Team Capabilities

Figure 2.4: Goleman's 1995 EI Model

Just as Bar-On drew on an array of non-cognitive capabilities, competencies, and skills, Goleman's model of EI is also a mixed model and it is characterized by five constructs. These constructs are: 1) knowing one's emotions (**Self Awareness**), 2) managing emotions (**Self Management**), 3) motivating oneself (**Self Regulation**), 4) recognizing emotions in others (**Social Awareness**), and 5) handling relationships (**Relationship Management**). Goleman included a set of emotional competencies within each construct of EI. Emotional competencies are not innate talents, but rather learned capabilities that must be practiced and developed to achieve outstanding performance. Goleman posits that individuals are born with a general emotional intelligence that determines their potential for learning emotional competencies. By focusing on a person's potential, Goleman's model breaks from the initial work of Salovey and Mayer who define emotional intelligence as a set of abilities.

Goleman's landmark article, published in Harvard Business Review (HBR), see [80], expanded on

TTI SUCCESS INSIGHTS®

the components described in [79]. His 1995 framework identified five components of emotional intelligence that comprised twenty-five competencies. Three dimensions - Self Awareness, Self Regulation, and Motivation - describe personal competencies (knowing and managing emotions in oneself). The final two dimensions - Empathy and Social Skills - describe social competencies (knowing and managing emotions in others). Figure 2.4 summarizes Goleman's 1995 model. The information in Figure 2.4 is presented as follows. Numerals 1 through 5 denote the construct being measured. This is followed by, in each case, a definition denoted by (a). The lowercase letters indicate the sub-constructs or competencies as mentioned earlier.

Based on a statistical analysis conducted by Boyatzis, Goleman's 1995 model was refined, collapsing the 25 competencies into 20 and the five domains into four: Self Awareness, Self Management, Social Awareness, and Relationship Management, see [31]. The domains and their related competencies are presented in Table 2.1.

Table 2.1: Goleman's 2000 EI Model

	Self (Personal Competence)	**Other (Social Competence)**
Recognition	**Self Awareness** · Emotional Self Awareness · Accurate Self Assessment · Self Confidence	**Social Awareness** · Empathy · Service Orientation · Organizational Awareness
Regulation	**Self Management** · Emotional Self Control · Trustworthiness · Conscientiousness · Adaptability · Achievement Drive · Initiative	**Relationship Management** · Developing Others · Influence · Communication · Conflict Management · Visionary Leadership · Catalyzing Change · Building Bonds · Teamwork and Collaboration

Notably absent in this most recent version of the Goleman model is the Motivation construct. While inclusion of Motivation in the 1995 model came about as a result of workplace literature studies, the removal was due to a combination of scale reliability findings by Boyatzis, et. al., presented in [31], and the incorporation of neurological findings. Analysts suggested the removal of Motivation as a construct while neurologists offer an explanation of the role Motivation plays in all four domains, especially Self Management.

The study of decision making pathways in [61] explains the roles of played by neural-networks that include the dorsolateral, ventromedial, and orbitofrontal sectors of the prefrontal cortex. Collura, et. al., see [48] and [49], added to this understanding by demonstrating gamma asymmetry in these key right and left side networks and their interaction with the amygdala and hippocampus.

Bonnstetter, et. al., in [21] and [23], as well as Collura, et. al., in [53], found that the greater the activity level in the left medial prefrontal cortex, the more positive the person's emotional state. This suggests that the ability to regulate negative affect appears to be the circuit between the amygdala and the prefrontal cortex, with a left activation resulting in a more positive response

and a right dominant flare leading to a more negative reaction. These findings help show that Motivation may be viewed as a major contributor to EI, but more research is likely required to expose the more specific role it plays.

2.4 A Comparison of the Salovey, Bar-On, and Goleman EI Models

Figure 2.5 compares and contrasts the three initial EI models. While all three models address the intersection of emotion and cognition, the underlying fundamental elements differ. The Mayer & Salovey model, presented in the first column of Figure 2.5, describes EI as a set of unique emotional abilities or skills. These abilities are presented as hierarchical levels with the most basic skills involving perception and appraisal of emotions. The second skill set involves assimilating these initial perceptions and allowing emotions to direct attention in context. This second level may best be described as the ability to apply appropriate emotional responses, but may lack a rationale for the response.

Three Competing Models of Emotional Intelligence*

Mayer & Salovey (1997)	Bar-On (1997)	Goleman (1995)
Overall Definition "Emotional intelligence is the set of abilities that account for how people's emotional perception and understanding vary in their accuracy. More formally, we define emotional intelligence as the ability to perceive and express emotion, assimilate emotion in thought, understand and reason with emotion and regulate emotion in the self and others". (Mayer & Salovey, 1997)	Overall definition "Emotional intelligence is an array of noncognitive capabilities, competencies, and skills that influence one's ability to succeed in coping with environmental demands and pressures". (Bar-On, 1997)	Overall definition "The abilities called here emotional intelligence, which include self control, zeal and persistence, and the ability to motivate oneself" (Goleman, 1995). "There is an old-fashioned word for the body of skills that emotional intelligence represents: Character" (Goleman, 1995).
Major Areas of skills and Specific Examples *Perception* and Expression of Emotion Identifying and expressing emotions in one's physical states, feelings, and thoughts. Identifying and expressing emotions in other people, artwork, language, etc.	**Major Areas of skills and Specific Examples** *Intrapersonal Skills* Emotional self-awareness Assertiveness Self-regard Self-Actualisation Independence *Interpersonal Skills* Interpersonal relationships	**Major Areas of skills and Specific Examples** *Knowing One's Emotions* Recognizing a feeling as it happens. Monitoring feelings from moment to moment *Management of Emotions* Handling feelings so they are appropriate.
Assimilating Emotion in Thought Emotions prioritize thinking in productive ways Emotions generated as aids to judgment and memory *Understanding and Analysing Emotions* Ability to label emotions, including complex emotions and simultaneous feelings Ability to understand relationships associated with shifts of emotions *Reflective Regulation of Emotion* Ability to stay open to feelings Ability to monitor and regulate emotions reflectively to promote emotional and intellectual growth (Mayer and Salovey, 1997)	Social responsibility Empathy *Adaptability Scales:* Problem solving Reality testing Flexibility *Stress Management Scales:* Stress tolerance Impulse control *General mood:* Happiness Optimism	Ability to soothe oneself Ability to shake off rampant anxiety, gloom, or irritability *Motivating Oneself* Marshalling emotions in the service of a goal Delaying gratification and shifting impulsiveness Being able to get into the "flow" state *Recognizing Emotions in Others* Empathic awareness Attunement to what others need or want *Handling Relationships* Skill in managing emotions in others Interacting smoothly with others
Model Type Ability	**Model Type** Mixed	**Model Type** Mixed

Source: Models of Emotional Intelligence by Mayer et al. (2000)

Figure 2.5: Comparison of 3 Initial EI Models

The third progression results in understanding and the ability to logically analyze emotion and not simply respond appropriately. The fourth and highest level involves the ability to manage and regulate the emotions of oneself and others. All four levels are directly tied to a set of mental abilities or skills that are representative of this continuum.

Column 2 (Bar-On) and column 3 (Goleman) in Figure 2.5 are referred to in the literature as mixed models due to their blending of mental abilities and non-cognitive capacities. While Mayer & Salovey's model has a focus on observable skills as an end product, Bar-On and Goleman broaden

TTI SUCCESS INSIGHTS®

their models to include the capacity or potential for achieving success rather than success itself. This subtle difference results in the Bar-On model having the skills of intra- and inter- personal, plus scales for adaptability, stress management, and general mood. In the case of Goleman, his 1995 mixed model includes knowing one's emotions, management of emotions, motivating oneself, recognizing emotions in others, and handling relationships. Goleman summarizes the mixed nature of this model by stating:

> There is an old-fashioned word for the body of skills that emotional intelligence represents: character, [79], p. 285.

As is the case with any evolving body of knowledge, the more one learns about a topic, the more complex the topic has a tendency to become and increasingly more unknowns are uncovered. The Goleman model may be considered a more "mature" theory and, as a result, it requires more complex assessment tools.

2.5 The Next Generation of EI Assessments

Based on the work of these three pioneer EI models, numerous assessments emerged. Some of the earliest include Catanzaro & Mearns, [40], Giuliano & Swinkels, [78], Bagby, et. al., [7, 8], Cooper & Sawaf, [57], Salovey, et. al., [152], Swinkels & Giuliano, [168], Bar-On, [14], Wang, et. al., [175], and Schutte, et. al., [160]. The initial reviews by Davies, et. al., [62], as well as Mayer, et. al., [119], questioned these assessments regarding discrimination, construct validity, and other measures of overall assessment validity. It is important to note that the reviews do not conclude that the tools do not work, only that far more rigorous design protocols need to be followed and more in-depth studies performed.

2.6 Summary

This section of the technical manual provides an overview of the historical development and theoretical foundations of Emotional Intelligence (EI), leading to the creation of the TTI Success Insights Emotional Quotient (EQ) assessment. It traces the evolution of EI from early psychological studies on emotions and cognition to the modern frameworks used in assessing emotional intelligence.

The history of EI begins with early research in the 1980s that explored the influence of emotions on thought and decision-making. The concept of EI was first formally defined by Salovey and Mayer (1990) as a form of social intelligence involving emotional awareness, regulation, and utilization. This work laid the foundation for multiple EI models, including:

- The Ability Model (Salovey & Mayer) - EI as a measurable cognitive ability with four branches: emotional perception, emotional facilitation of thought, emotional understanding, and emotional management.
- The Trait Model (Bar-On) - Introduced the term Emotional Quotient (EQ), defining EI as a blend of emotional and social skills that influence life success.
- The Competency Model (Goleman) - Focused on workplace applications, linking EI to leadership and professional success through competencies such as self-awareness, self-regulation, motivation, empathy, and social skills.

Each model presents a unique perspective on EI, influencing various assessment tools and applications. The document also highlights the ongoing refinement of EI assessment methodologies, emphasizing the importance of validity, reliability, and scientific rigor.

Finally, the section discusses the next generation of EI assessments, recognizing the increasing complexity of measuring EI and the need for continuous research to refine assessment tools. It underscores the importance of EI in both personal and professional development, setting the stage for more advanced and comprehensive approaches to understanding emotional intelligence.

3 Our Approach to the Emotional Quotient Assessment

In order for a complete discussion on the TTI Success Insights EQ model to occur, we must first outline the evolution of the Goleman model over the years. In the previous section on the history of emotional intelligence, we outlined the 1995 and 2000 versions of Goleman's model. The story is much broader and we present a more detailed documentation of this evolution in this section. We finish the section by detailing the final model employed by TTI Success Insights as well as a brief discussion of our plans for the future.

3.1 Goleman's 1995 EQ Model

Goleman's 1995 model is presented in Figure 2.4 in Section 2.3 above. However, that presentation does not go in depth into how Goleman defines the sub-constructs used to define the independent constructs in this version of the model. We present both the constructs and sub-constructs, complete with definitions here.

1. Self Awareness - the ability to recognize and understand your moods, emotions, and drives, as well as their effect on others.
   (a) Emotional Self Awareness - the ability to recognize and understand one's own emotions and their effects. This involves being conscious of one's emotional state and how it influences one's behavior and decision-making. Emotional self awareness is crucial for self-reflection and understanding one's strengths, weaknesses, and triggers.
   (b) Accurate Self Assessment - the ability to have a realistic understanding of one's strengths and limitations. Individuals who are adept at this can accept constructive criticism, recognize areas for personal growth, and are aware of their skills and areas needing improvement. This self assessment leads to great self confidence and personal development.
   (c) Self Confidence - a strong sense of one's self worth and capabilities. Individuals with high self confidence trust their abilities, make decisions assertively, and are willing to take on new challenges. They are comfortable speaking up, expressing opinions, and defending their viewpoints.
2. Self Regulation - the ability to control or redirect disruptive impulses and moods. The propensity to suspend judgment - to think before acting.
   (a) Self Control - the ability to manage disruptive emotions and impulses effectively. This includes staying calm and clear-headed in stressful situations and not making hasty decisions or taking hasty actions.
   (b) Trustworthiness - maintaining honesty and integrity. This involves being consistent in words and actions and upholding ethical standards, which builds trust in professional and personal relationships.
   (c) Conscientiousness - being responsible, dependable, and accountable. Conscientious individuals are organized, careful, and reliable in managing their work and commitments.
   (d) Adaptability - being flexible in handling change and ambiguity. This aspect involves adjusting to new information, challenges, or unexpected events without becoming overly stressed or disoriented.

(e) Innovation - being open to new ideas, approaches, and information. Individuals who demonstrate innovation are willing to think outside the box, take calculated risks, and approach situations creatively.

3. Motivation - a passion to work for reasons that go beyond money or status. A propensity to pursue goals with energy and persistence.

   (a) Achievement Drive - a strong desire to meet or exceed a standard of excellence. This includes setting challenging goals, taking calculated risks, and continuously seeking ways to improve and measure performance against high standards.

   (b) Commitment - aligning with the goals of a group, organization, or mission. Individuals with high motivation are dedicated to their objectives and are willing to make sacrifices or go the extra mile to achieve them.

   (c) Initiative - the readiness to act and seize opportunities. This involves being proactive rather than reactive, taking steps to solve problems or make progress even when not specifically directed or asked to do so.

   (d) Optimism - maintaining a positive outlook, even in the face of challenges and setbacks. Optimistic individuals view setbacks as learning opportunities rather than as failures, and they persist in pursuing their goals despite obstacles.

4. Empathy (Social Awareness) - the ability to understand the emotional makeup of other people. Skill in treating people according to their emotional reactions.

   (a) Understanding Others - the ability to sense and understand the emotions, perspectives, and needs of others. Empathy involves recognizing emotional cues, understanding others' feelings, and taking an active interest in their concerns. It is fundamental to building relationships, managing teams, and leading others. Note, we are using a slightly modified definition of empathy as a sub-construct that Goleman provides in his 1995 and 1998 writings on the topic.

   (b) Developing Others - sensing others' development needs and bolstering their abilities. People with this competency acknowledge and reward others' strengths and accomplishments, offer useful feedback, and identify others' needs for further growth. They also mentor, provide timely coaching, and offer assignments that challenge and foster an individual's skills.

   (c) Service Orientation - the ability to recognize and meet the needs of clients or customers. This construct involves actively seeking ways to help others and understand their needs, which is crucial for customer service roles, leadership, and any situations where one is responsible for providing support or guidance to others.

   (d) Leveraging Diversity - cultivating opportunities through different kinds of people. People with this competency respect and relate well to people from varied backgrounds, understand diverse worldviews, and are sensitive to group differences. They also see diversity as opportunity, help to create an environment where diverse peoples can thrive, and challenge bias and intolerance.

   (e) Political Awareness - reading the social and political current. People with this competency accurately read key power relationships, detect crucial social networks, and understand the forces that shape views and actions of clients, customers, or competitors. They also accurately read organizational and external realities.

5. Social Skills (Social Regulation) - proficiency in managing relationships and building networks. An ability to find common ground and build rapport.

   (a) Influence - the ability to impact others' actions, decisions, and attitudes.

   (b) Communication - effectively transmitting information and listening actively.

   (c) Conflict Management - the skills needed to resolve disagreements.

   (d) Leadership - providing guidance and motivating others towards a common goal.

   (e) Change Catalyst - initiating, managing, and leading change in a positive direction.

   (f) Building Bonds - creating and maintaining a network of relationships.

   (g) Collaboration and Cooperation - working with others toward shared goals.

   (h) Team Capabilities - creating group synergy in pursuing collective goals.

3.2 Goleman's 2000 EQ Model

The basic model is largely the same with some modifications in naming convention and construct or sub-construct definitions. The exception is that Goleman has removed Motivation as an independent construct.

1. Self Awareness - knowing one's internal states, preferences, resources, and intuitions.

   (a) Emotional Self Awareness - the ability to recognize and understand one's own emotions and their effects. This involves being conscious of one's emotional state and how it influences one's behavior and decision-making. Emotional self awareness is crucial for self-reflection and understanding one's strengths, weaknesses, and triggers.

   (b) Accurate Self Assessment - the ability to have a realistic understanding of one's strengths and limitations. Individuals who are adept at this can accept constructive criticism, recognize areas for personal growth, and are aware of their skills and ares needing improvement. This self assessment leads to great self confidence and personal development.

   (c) Self Confidence - a strong sense of one's self worth and capabilities. Individuals with high self confidence trust their abilities, make decisions assertively, and are willing to take on new challenges. They are comfortable speaking up, expression opinions, and defending their viewpoints.

2. Self Management - the ability to manage and control one's internal states, impulses, and resources.

   (a) Emotional Self Control - keeping disruptive emotions and impulses in check. This includes staying calm and clear-headed in stressful situations and not making hasty decisions or actions.

   (b) Trustworthiness - maintaining standards of honesty and integrity. This involves being consistent in words and actions and upholding ethical standards, which builds trust in professional and personal relationships.

   (c) Conscientiousness - taking responsibility for personal performance. Conscientious individuals are organized, careful, and reliable in managing their work and commitments.

(d) Adaptability - being flexible in handling change. This aspect involves adjusting to new information, challenges, or unexpected events without becoming overly stressed or disoriented.

(e) Achievement Drive - striving to improve or meet a standard of excellence. This includes setting challenging goals, taking calculated risks, and continuously seeking ways to improve and measure performance against high standards.

(f) Initiative - readiness to act on opportunities.

3. Motivation - emotional tendencies that guide or facilitate reaching goals.

   (a) Commitment - aligning with the goals of the group or organization.

   (b) Optimism - persistence in pursuing goals despite obstacles and setbacks

   Note that we include the Motivation construct for informative purposes. As the alert reader has noticed, Achievement Drive and Initiative have moved to Self Management in the updated model. With some reservation, we would also argue that it may be the case that with the above definition for Commitment, this sub-construct may be well suited to fall under Organizational Awareness under the construct Social Awareness that appears next. Note also that the sub-construct Developing Others from the 2000 Empathy construct has moved to the Relationship Management construct.

4. Social Awareness - the ability to understand the emotional makeup of other people. Skill in treating people according to their emotional reactions. Reading people and groups accurately.

   (a) Empathy - the ability to sense and understand the emotions, perspectives, and needs of others. Empathy involves recognizing emotional cues, understanding others' feelings, and taking an active interest in their concerns. It is fundamental to building relationships, managing teams, and leading effectively.

   (b) Service Orientation - the ability to recognize and meet the needs of clients or customers. This construct involves actively seeking ways to help others and understand their needs, which is crucial for customer service roles, leadership, and any situations where one is responsible for providing support or guidance to others.

   (c) Organizational Awareness - the ability to read the currents, decision networks, and politics within an organization. This involves understanding the dynamics at play in a group, such as recognizing power relationships, informal networks, and organizational culture. Organizational awareness is essential for navigating complex social structures and influencing others.

5. Relationship Management (Social Regulation) - proficiency in managing relationships and building networks. An ability to find common ground and build rapport. Inducing desirable responses in others.

   (a) Developing Others - sensing others' development needs and bolstering their abilities.

   (b) Influence - the ability to impact others' actions, decisions, and attitudes.

   (c) Communication - effectively transmitting information and listening actively.

   (d) Conflict Management - the skills needed to resolve disagreements.

(e) Visionary Leadership - drawing on a range of personal skills to inspire others to work together toward common goals.

(f) Catalyzing Change - initiating, managing, and leading change in a positive direction.

(g) Building Bonds - creating and maintaining a network of relationships.

(h) Teamwork and Collaboration - using one's personal skills to help establish an environment that fosters collaboration, as well as the ability to facilitate that collaboration into an effective team setting.

3.3 TTI Success Insights' Current EQ Model

The current iteration of the TTI Success Insights Emotional Intelligence model is essentially the original Goleman 1995 model, see Section 3.1. As time has evolved, Goleman discovered that Motivation was not an independent construct of emotional intelligence and has taken some of the sub-constructs he outlines in his 2000 model, see Section 3.2, and moved them to be part of other independent constructs. Due to the overwhelming popularity and use of Motivation in our network by our clients and our clients' clients, we have maintained Motivation as a score we present. The current iteration does, in fact, still view Motivation as an EI construct. However, as we present in Section 3.4, our long term vision is to remove it as a construct.

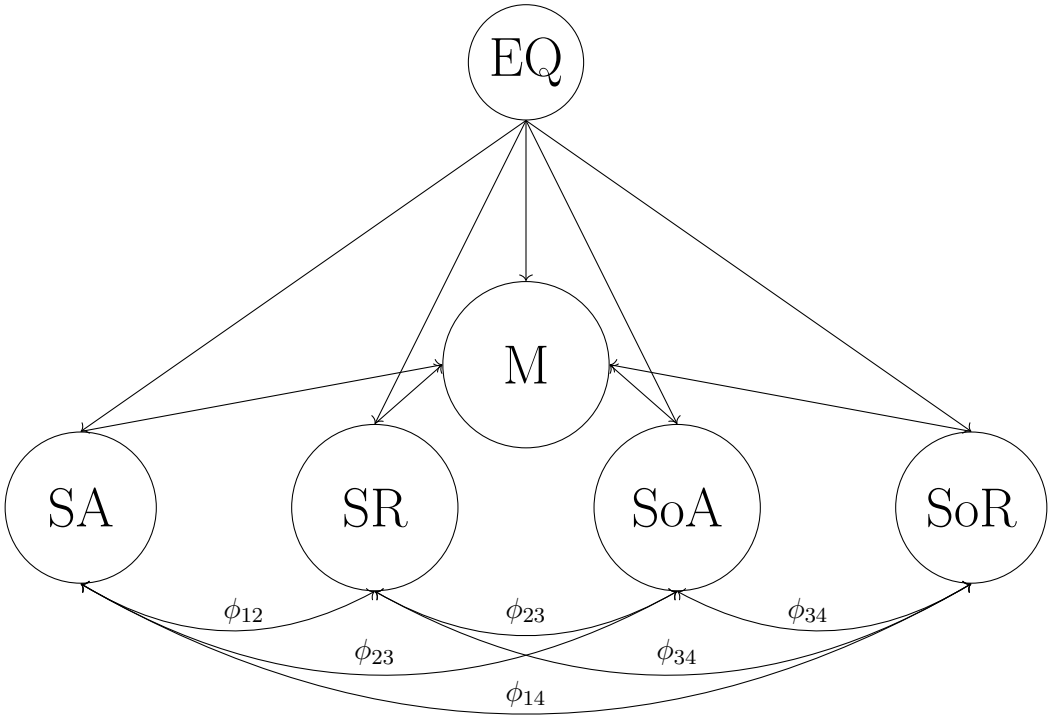


Figure 3.1: TTI Success Insights Current Emotional Intelligence Model

To the reader concerned about losing the ability to have Motivation as part of the report and proceeding discussions, we intend to present a motivation score by independently scoring the relevant sub-constructs. As an example, Achievement Drive and Initiative in the 1995 Goleman model are moved to the Self Management construct in his 2000 model. In other words, Motivation has

been removed, but these two concepts are still being measured. Commitment, defined as aligning with the goals of the group or organization, may be thought of as part of Service Orientation or Organizational Awareness in the Social Awareness construct. Optimism may or may not still be part of this model, and decisions regarding how to deal with this piece are ongoing.

Regardless of the outcome of the discussions surrounding the Motivation score, the goal of TTI Success Insights is to move toward a full implementation of the 2000 Goleman model as outlined and defined in the next section. This does mean an increase in the number of questionnaire items which may concern some. However, it is noted that the Bar-On ESI model questionnaire, one of our leading competitors in this space, has at least 132 items.

3.4 TTI Success Insights Ideal EQ Model

The TTI Success Insights Emotional Intelligence (EQ) model is comprised of four main constructs. The layout is visually displayed in Figure 3.2. The abbreviations used in the aforementioned figure are as follows. SA is short for Self Awareness which is outlined with its sub-constructs in Section 3.4.1 below. SR represents the construct Self Regulation which is defined, along with its sub-constructs, in Section 3.4.2. SoA represents the construct Social Awareness which is defined, along with its sub-constructs, in Section 3.4.4. Finally, SoR represents the construct Social Regulation which is defined, along with its sub-constructs, in Section 3.4.5.

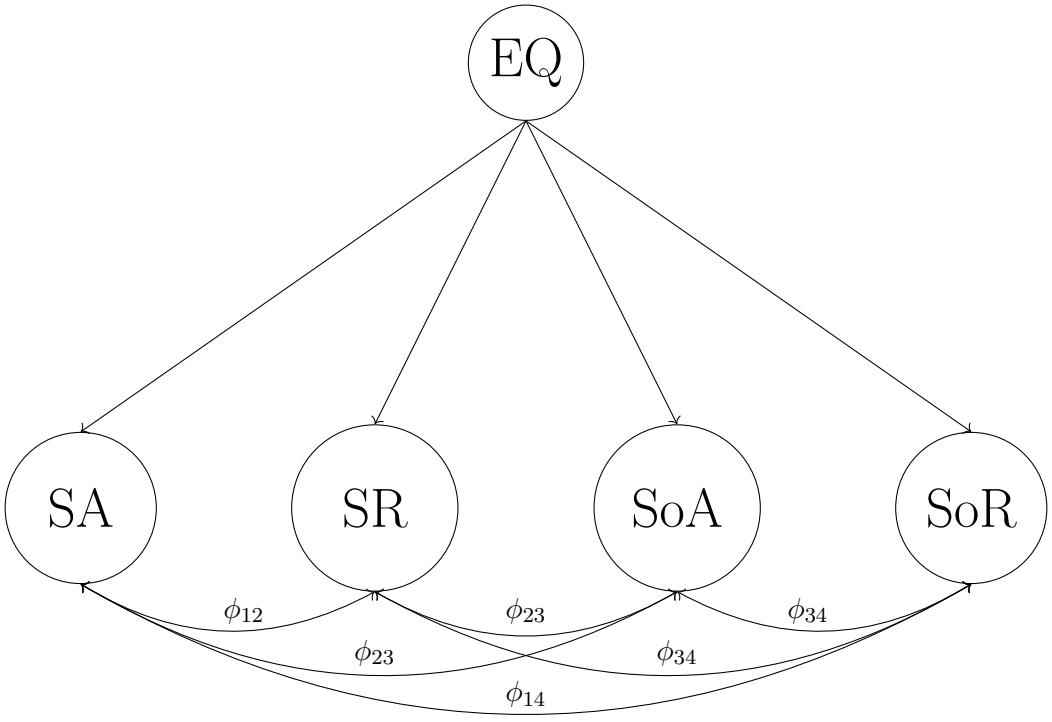


Figure 3.2: TTI Success Insights Emotional Intelligence Model

Note that the symbols ϕ_{ij} represent the correlation between the i^{th} and j^{th} constructs.

3.4.1 Self Awareness

Definition 3.1. ***Self Awareness*** *is defined as the ability to recognize and understand one's own emotions, thoughts, and behaviors, as well as how they impact others.*

Self Awareness serves as a foundation for personal growth and effective interpersonal relationships, as well as providing a critical component of overall emotional intelligence. Self Awareness is comprised of three basic concepts: Emotional Self Awareness (ESA), Accurate Self Assessment (ASA), and Self Confidence (SC). This relationship is depicted in Figure 3.3. The symbols ψ_{ij} represent the intra-construct correlations between the i^{th} and j^{th} sub-constructs.

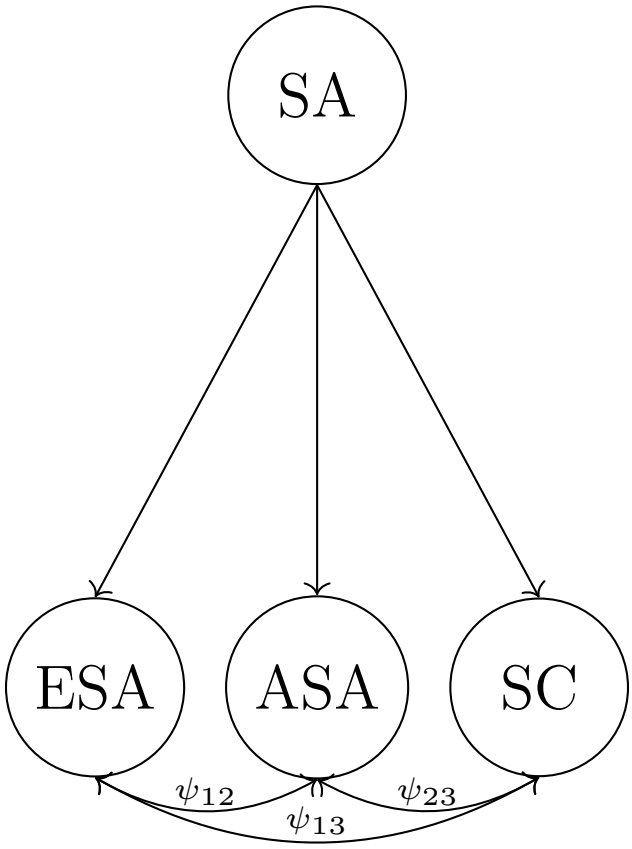


Figure 3.3: TTI Success Insights Self Awareness Construct

Definition 3.2. ***Emotional Self Awareness*** *(ESA) is defined as the ability to recognize and understand one's own emotions in real-time, as well as the implications of those emotions for thoughts and behaviors.*

Emotional Self Awareness is a critical component of overall self-awareness and plays a key role in supporting healthy emotional regulation and developing strong interpersonal effectiveness. The key components of this sub-construct are graphically represented in Figure 3.4 and include:

1. Identification of Emotions (IE) - accurately recognizing and labeling one's own emotions as they arise, whether they are positive or negative.

TTI SUCCESS INSIGHTS®

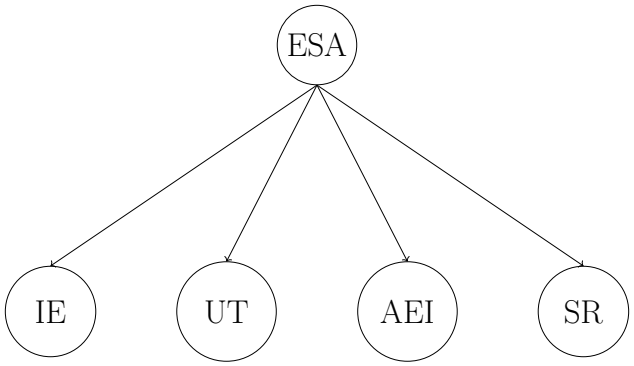


Figure 3.4: TTI Success Insights Emotional Self Awareness Sub-construct

2. Understanding Triggers (UT) - knowing what specific situations or events elicit certain emotional responses, which helps in anticipating and managing reactions.

3. Awareness of Emotional Impact (AEI) - understanding how one's emotions influence thoughts, decisions, and interactions with others, leading to more intentional behavior. This includes being receptive to others' observations about one's emotional behavior and using that feedback to enhance understanding and self-regulation

4. Self-Reflection (SR) - engaging in introspective practices to assess how emotions align with personal values and goals, fostering personal growth.

Overall, Emotional Self Awareness is about having a clear understanding of one's emotional landscape, which enables better emotional management, effective communication, and healthier relationships.

Definition 3.3. ***Accurate Self Assessment*** *is defined as the ability to evaluate one's own strengths, limitations, and emotional capabilities realistically.*

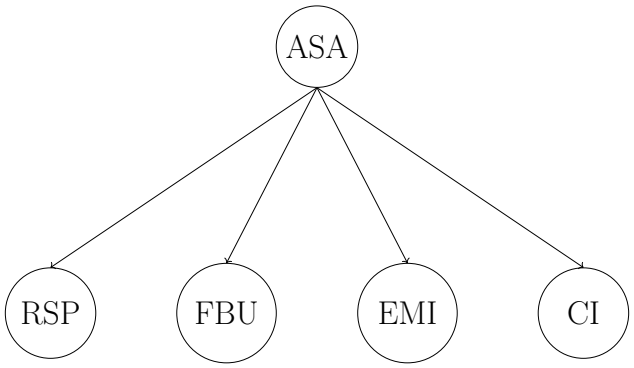


Figure 3.5: TTI Success Insights Accurate Self Assessment Sub-construct

Accurate Self Assessment involves a clear understanding of one's skills and limitations, which helps foster personal and professional growth. The key components of this sub-construct are graphically represented in Figure 3.5 and include:

1. Realistic Self-Perception (RSP) - recognizing one's abilities and limitations without overestimating or underestimating oneself, leading to a balanced self-image.
2. Feedback Utilization (FBU) - actively seeking and valuing feedback from others to gain insights into one's performance and emotional impact on others.
3. Emotional Insight (EMI) - understanding how one's emotional state influences actions and decisions, contributing to better self-management.
4. Continual Improvement (CI) - committing to ongoing learning and development based on self-assessment, adapting to new challenges and opportunities.

Overall, Accurate Self Assessment in emotional intelligence is about having a clear and honest understanding of oneself, which enables effective decision-making, personal growth, and improved interpersonal relationships.

Definition 3.4. ***Self Confidence*** *is defined as the belief in one's own abilities, qualities, and judgment.*

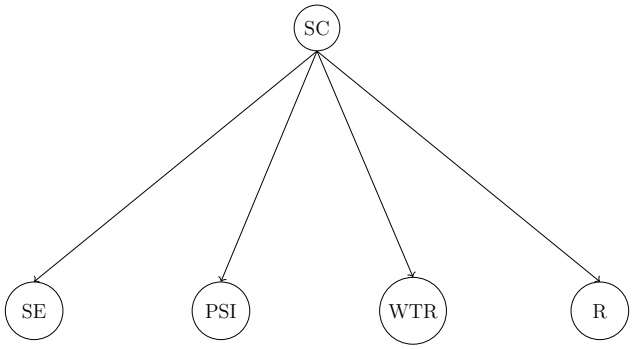


Figure 3.6: TTI Success Insights Self Confidence Sub-construct

Self Confidence encompasses a positive self-perception and assurance in one's capacity to handle challenges and achieve goals. The key components of this sub-construct are graphically represented in Figure 3.6 and include:

1. Self-Efficacy (SE) - the belief in one's ability to successfully perform tasks and overcome obstacles, which fosters resilience and motivation.
2. Positive Self-Image (PSI) - maintaining a healthy view of oneself, including recognizing strengths while accepting limitations without excessive self-criticism.
3. Willingness to Take Risks (WTR) - being open to new challenges and opportunities, showing a readiness to step outside of comfort zones based on belief in one's capabilities.
4. Resilience (R) - demonstrating the ability to recover from setbacks and learn from failures, reinforcing belief in one's capacity to succeed.

Note 3.5. *We note here that the use of the term* ***Resilience*** *is not intended to be in the sense of the use of the term as defined by the American Psychological Association (APA). The APA*

TTI SUCCESS INSIGHTS®

defines Resilience in the psychometric sense as the process and outcome of successfully adapting to difficult or challenging life experiences, especially through mental, emotional, and behavioral flexibility and adjustment to external and internal demands. Our use of the term is more consistent with the English language definition of the word, the capacity to withstand or to recover quickly from difficulties.

Overall, Self Confidence in emotional intelligence is about having a strong belief in one's abilities and potential, which enhances performance, fosters positive interactions, and contributes to overall emotional well-being.

3.4.2 Self Regulation

Definition 3.6. ***Self Regulation*** *is defined as the ability to manage one's emotions, thoughts, and behaviors effectively, especially in challenging situations.*

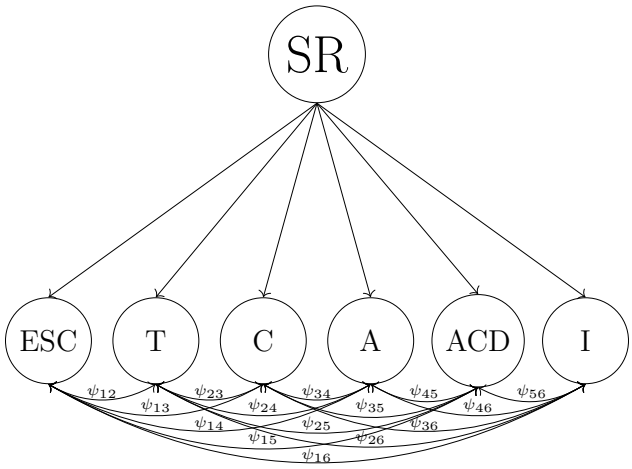


Figure 3.7: TTI Success Insights Self Regulation Construct

Self Regulation plays a crucial role in maintaining emotional balance and making thoughtful decisions as well as providing a critical component of overall emotional intelligence. Self Regulation is comprised of six key concepts: Emotional Self Control (ESC), Trustworthiness (T), Conscientiousness (C), Adaptability (A), Achievement Drive (AD), and Initiative (I). This relationship is depicted in Figure 3.7. The symbols ψ_{ij} represent the intra-construct correlations between the i^{th} and j^{th} sub-constructs.

Definition 3.7. ***Emotional Self Control*** *is defined as the ability to manage and regulate one's emotions in a way that promotes thoughtful decision-making and appropriate responses, particularly in challenging or stressful situations.*

Emotional Self Control is a critical part of effective self regulation. The key components of this sub-construct are graphically represented in Figure 3.8 and include:

1. Emotional Awareness (EA) - recognizing and understanding one's own emotional triggers and responses, which is crucial for effective self-control.

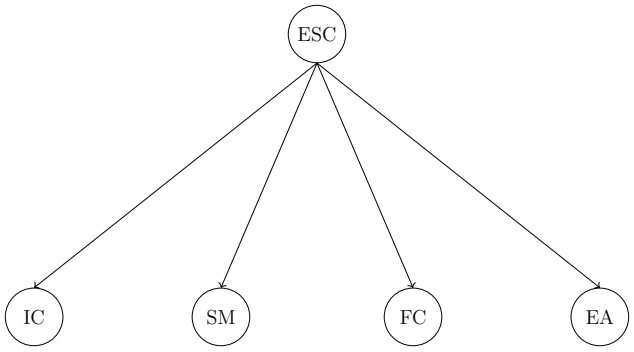


Figure 3.8: TTI Success Insights Emotional Self Control Sub-construct

2. Impulse Control (IC) - the ability to resist immediate emotional reactions or urges, allowing for more deliberate and considered actions.

3. Stress Management (SM) - employing techniques to remain calm and composed under pressure, facilitating better emotional regulation and decision-making.

4. Focus and Concentration (FC) - maintaining attention on tasks and goals, even when faced with distractions or emotional upheaval, to enhance productivity.

Overall, Emotional Self Control in emotional intelligence is about mastering one's emotional landscape to foster resilience, maintain focus, and engage positively with others, even in difficult situations.

Definition 3.8. ***Trustworthiness*** *is defined as the ability to inspire confidence in others through integrity, reliability, and ethical behavior.*

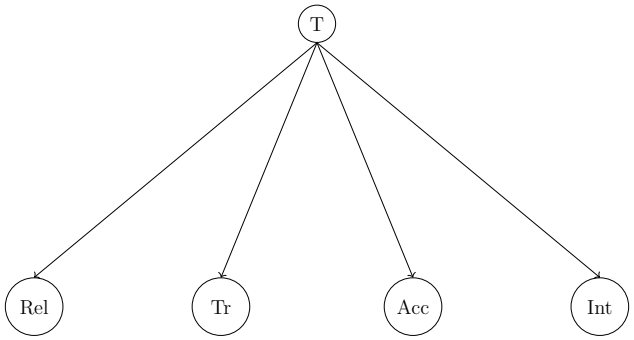


Figure 3.9: TTI Success Insights Trustworthiness Sub-construct

Trustworthiness is essential for building strong relationships and fostering a positive environment. The key components of this sub-construct are graphically represented in Figure 3.9 and include:

1. Integrity - adhering to moral and ethical principles, ensuring consistency between words and actions, which fosters credibility.

2. Reliability - being dependable and following through on commitments, which reinforces trust among colleagues and team members.

TTI SUCCESS INSIGHTS®

3. Transparency - openly sharing information and decisions, allowing others to understand motivations and processes, which enhances trust.

4. Accountability - taking responsibility for one's actions and decisions, acknowledging mistakes, and learning from them, which demonstrates honesty and integrity.

Overall, Trustworthiness in emotional intelligence is about cultivating an environment where individuals feel secure and valued, leading to stronger collaboration and deeper connections.

Definition 3.9. ***Conscientiousness*** *is defined as the ability to be diligent, responsible, and reliable in one's actions and decisions.*

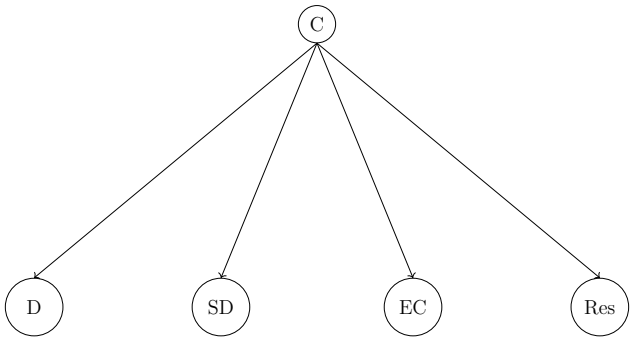


Figure 3.10: TTI Success Insights Conscientiousness Sub-construct

Conscientiousness involves a strong sense of duty and the ability to follow through on commitments. The key components of this sub-construct are graphically represented in Figure 3.10 and include:

1. Responsibility - taking ownership of one's tasks and obligations, ensuring that duties are met with care and consideration.

2. Dependability - consistently delivering on promises and being reliable, which fosters trust and confidence among colleagues and team members.

3. Self-Discipline - managing impulses and distractions effectively, allowing for sustained effort and productivity.

4. Ethical Consideration - upholding high standards of integrity and ethics in decision-making and actions, which reinforces reliability and trustworthiness.

Overall, Conscientiousness in emotional intelligence is about being responsible and proactive, contributing positively to team dynamics and fostering a culture of accountability and excellence.

Definition 3.10. ***Adaptability*** *is defined as the ability to adjust one's thoughts, emotions, and behaviors in response to changing circumstances or new information.*

Adaptability plays a crucial role in navigating uncertainty and maintaining effectiveness in dynamic environments. The key components of this sub-construct are graphically represented in Figure 3.11 and include:

1. Flexibility - the willingness to change plans or approaches when faced with new challenges or feedback, allowing for innovative solutions.

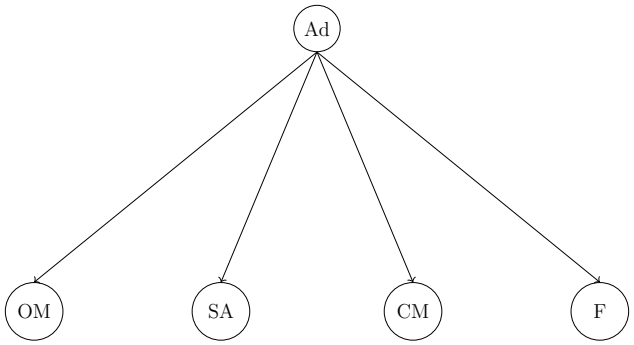


Figure 3.11: TTI Success Insights Adaptability Sub-construct

2. Open-Mindedness - being receptive to new ideas, perspectives, and experiences, which enhances problem-solving and creativity.
3. Situational Awareness - understanding the context and dynamics of a situation, allowing for timely adjustments to strategies and behaviors.
4. Collaborative Mindset - working effectively with others in a flexible manner, fostering teamwork and communication to adapt to group needs.

Overall, Adaptability in emotional intelligence is about being responsive and open to change, facilitating personal growth, effective collaboration, and resilience in the face of challenges.

Definition 3.11. ***Achievement Drive*** *is defined as the intrinsic motivation and determination to pursue and attain personal and professional goals.*

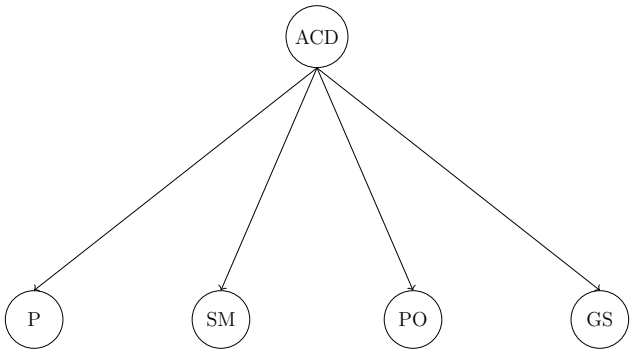


Figure 3.12: TTI Success Insights Achievement Drive Sub-construct

Achievement Drive involves a strong commitment to excellence and a desire for continual improvement. The key components of this sub-construct are graphically represented in Figure 3.12 and include:

1. Goal Setting - establishing clear, specific, and challenging objectives that drive focus and direction in one's efforts.
2. Persistence - demonstrating resilience and determination in the face of obstacles or setbacks, maintaining effort until goals are achieved.

3. Self-Motivation - finding internal sources of motivation that propel one to take initiative and pursue goals without external prompting.

4. Performance Orientation - striving for high standards of quality and excellence in one's work, continually seeking to improve and excel.

Overall, Achievement Drive in emotional intelligence is about fostering a proactive and results-oriented mindset that encourages continual growth, excellence, and a commitment to personal and professional success.

Definition 3.12. ***Initiative*** *is defined as the ability to take proactive steps, seize opportunities, and act on ideas without needing external prompts.*

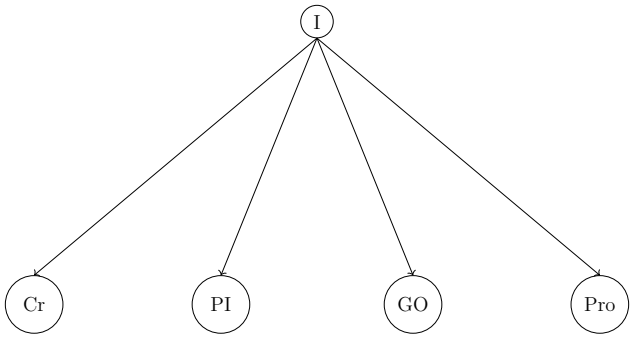


Figure 3.13: TTI Success Insights Initiative Drive Sub-construct

Initiative reflects a willingness to make things happen and drive positive change. The key components of this sub-construct are graphically represented in Figure 3.13 and include:

1. Proactivity - taking the lead in identifying challenges or opportunities and acting on them before being asked or directed.

2. Creativity - generating innovative ideas and approaches, thinking outside the box to find effective solutions to challenges.

3. Problem Identification - recognizing issues or areas for improvement and taking the initiative to address them constructively.

4. Goal Orientation - setting personal and professional goals and actively pursuing them, demonstrating a focus on achievement and progress.

Overall, Initiative in emotional intelligence is about being proactive and forward-thinking, driving positive actions and outcomes while fostering a culture of engagement and innovation.

3.4.3 Motivation

As noted in Section 3.3, it is the intent of TTI Success Insights to follow the Goleman 2005 EI model in as much as possible. To that extent, Motivation will no longer be considered an independent construct. However, we will attempt to continue to provide a motivation score in the report based on the fact that while Goleman removed Motivation, he did not remove all the sub-constructs being measured. As noted in Section 3.3, several of the Motivation sub-constructs are part of Self Regulation and Social Awareness in the updated model.

3.4.4 Social Awareness

Definition 3.13. *Social Awareness is defined as the ability to recognize and understand emotions, needs, and dynamics of others in social interactions.*

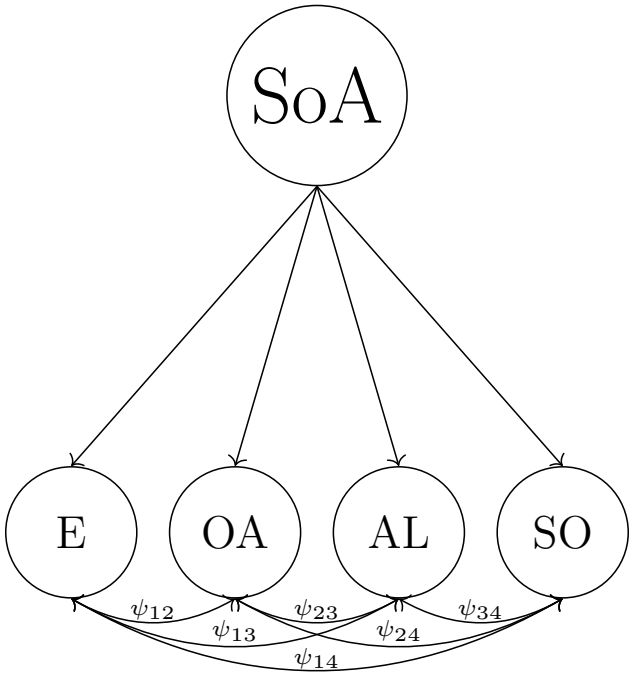


Figure 3.14: TTI Success Insights Social Awareness Construct

Social Awareness plays a crucial role in building strong relationships and effective communication. It is comprised of three main components: Empathy (EM), Service Orientation (SO), and Organizational Awareness (OA). This relationship is depicted in Figure 3.14. The symbols ψ_{ij} represent the intra-construct correlations between the i^{th} and j^{th} sub-constructs.

Definition 3.14. ***Empathy*** *is defined as the ability to understand, recognize, and share the feelings and perspectives of others.*

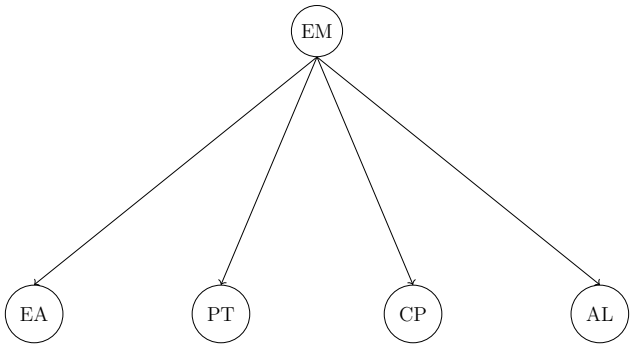


Figure 3.15: TTI Success Insights Empathy Sub-construct

TTI SUCCESS INSIGHTS®

Empathy plays a vital role in building strong relationships and fostering effective communication. The key components of this sub-construct are graphically represented in Figure 3.15 and include:

1. Emotional Awareness (EA) - the ability to recognize and understand one's own emotions and those of others, facilitating the ability to accurately identify and interpret the emotions of others through verbal and non-verbal cues, as well as fostering deeper connections and sense of understanding.
2. Perspective-Taking (PT) - understanding and appreciating the viewpoints and experiences of others, which helps in relating to their feelings and reactions.
3. Compassion (CP) - demonstrating genuine care and concern for others' well-being, motivating supportive actions and responses.
4. Active Listening (AL) - engaging fully with others by listening attentively, asking clarifying questions, and reflecting back feelings, which validates their emotions.

Overall, Empathy in emotional intelligence is about fostering connections through understanding and compassion, enhancing interpersonal relationships, and facilitating effective collaboration and support.

Definition 3.15. ***Service Orientation*** *is defined as the ability to recognize and meet the needs of others, focusing on providing quality support and enhancing customer or team satisfaction.*

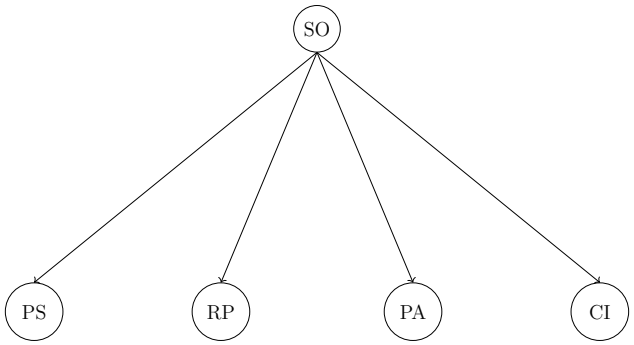


Figure 3.16: TTI Success Insights Service Orientation Sub-construct

Service Orientation plays a crucial role in building relationships and enhancing collaboration and teamwork. The key components of this sub-construct are graphically represented in Figure 3.16 and include:

1. Problem-Solving Skills (PSS) - identifying solutions to challenges faced by others, demonstrating a proactive approach to meeting their needs effectively.
2. Responsiveness (RP) - being quick to respond to inquiries or issues, showing commitment to service and creating a positive experience for others.
3. Positive Attitude (PA) - maintaining a friendly and approachable demeanor, which encourages open communication and fosters a supportive environment.

4. Commitment to Improvement (CI) - seeking feedback and continually looking for ways to enhance service quality, demonstrating a dedication to meeting and exceeding expectations.

Overall, Service Orientation in emotional intelligence is about prioritizing the needs of others, fostering positive interactions, and creating an environment where support and satisfaction are paramount.

Definition 3.16. ***Organizational Awareness*** *is defined as the ability to understand the dynamics, culture, and emotional landscape of an organization.*

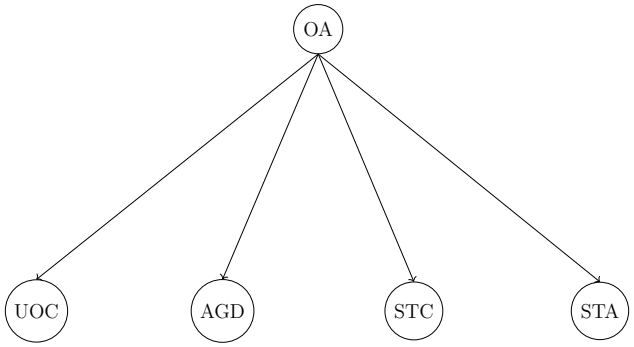


Figure 3.17: TTI Success Insights Service Orientation Sub-construct

Organizational Awareness enables individuals to navigate the social complexities within the workplace effectively. The key components of this sub-construct are graphically represented in Figure 3.17 and include:

1. Understanding Organizational Culture (UOC) - recognizing the values, beliefs, and behaviors that shape the organization, which informs how to interact appropriately within that context.
2. Awareness of Group Dynamics (AGD) - identifying roles, relationships, and power structures within teams, helping to anticipate how these dynamics influence interactions and decision-making.
3. Sensitivity to Change (STC) - recognizing signs of change within the organization, including shifts in morale, attitudes, and responses to new initiatives, allowing for timely and effective adjustments.
4. Stakeholder Awareness (STA) - identifying and understanding the needs, motivations, and concerns of various stakeholders, facilitating effective collaboration and engagement.

Overall, Organizational Awareness in emotional intelligence is about navigating the complexities of an organization's social environment, fostering effective communication and collaboration, and contributing positively to the organizational culture.

3.4.5 Social Regulation

Definition 3.17. ***Social Regulation*** *is defined as the ability to effectively manage interactions and relationships with others.*

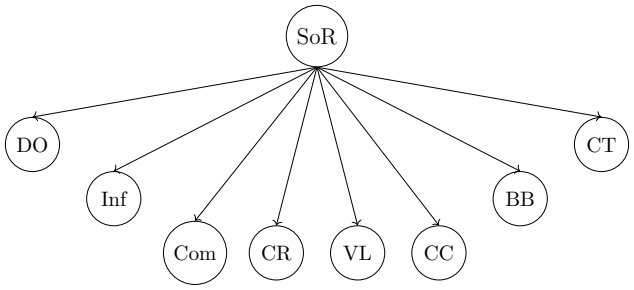


Figure 3.18: TTI Success Insights Social Regulation Construct

Social Regulation is crucial to proficiency in managing relationships and navigating social complexities, fostering collaboration, and building strong networks. Individuals with high levels of Social Regulation are adept at understanding the emotional needs of others, maintaining healthy boundaries, and resolving conflicts in a constructive manner. This construct highlights the importance of creating and sustaining positive interpersonal connections, which is essential for teamwork, leadership, and overall organizational success. Ultimately, managing relationships reflects a persons' capability to cultivate trust, cooperation, and emotional support within various contexts. Social Regulation is comprised of eight critical concepts: Developing Others (DO), Influence (Inf), Communication (Com), Conflict Resolution (CR), Visionary Leadership (VL), Catalyzing Change (CC), Building Bonds (BB), and Collaboration and Teamwork (CT). This relationship is depicted in Figure 3.18. The symbols ψ_{ij} are omitted to keep the figure less cluttered. Similar relationships exist as those depicted in Figure 3.7 and elsewhere in the above sections.

Definition 3.18. ***Developing Others*** *is defined as the ability to foster the growth and potential of individuals within a team or organization through emotional awareness and supportive interpersonal skills.*

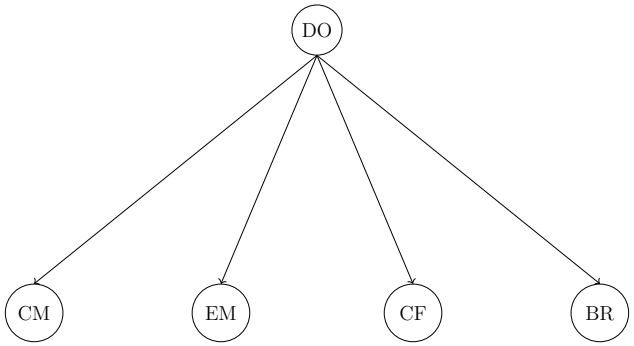


Figure 3.19: TTI Success Insights Developing Others Sub-construct

Developing Others is a critical component of effectively building and leading individuals and teams. The key components of this sub-construct are graphically represented in Figure 3.19 and include:

1. Coaching and Mentoring - the ability to provide guidance, feedback, and support tailored to individual needs, helping others enhance their skills and confidence.

2. Encouragement and Motivation - inspiring others to pursue their goals and take on challenges, fostering a positive mindset toward growth and learning.

TTI SUCCESS INSIGHTS®

3. Constructive Feedback - providing clear, specific, and actionable feedback that helps individuals improve while maintaining their dignity and confidence.

4. Building Relationships - establishing trust and rapport with team members, creating a supportive atmosphere conducive to personal and professional development.

Overall, this sub-construct highlights the role of emotional intelligence in nurturing talent and fostering a culture of continual learning, ultimately enhancing team effectiveness and organizational success.

Definition 3.19. ***Influence*** *is defined as the ability to effectively affect or change the attitudes, beliefs, and behaviors of others through understanding and managing emotions.*

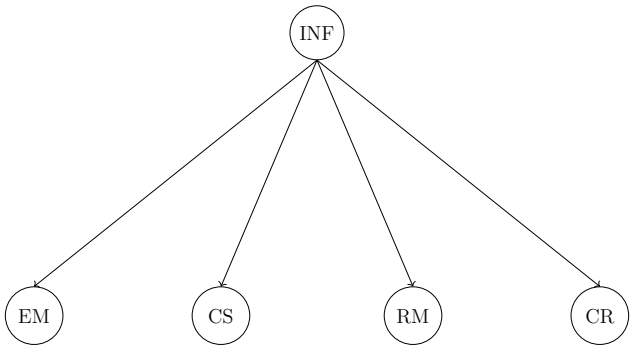


Figure 3.20: TTI Success Insights Influence Sub-construct

Influence is critical to building and leading individuals and teams. The key components of this sub-construct are graphically represented in Figure 3.20 and include:

1. Empathy - understanding others' perspectives and feelings, helping to build rapport and trust, making it easier to influence them.

2. Communication Skills - articulating ideas clearly and persuasively, using both verbal and non-verbal cues to resonate with others emotionally.

3. Relationship Management - building and maintaining strong interpersonal connections that facilitate influence, encouraging collaboration and cooperation.

4. Conflict Resolution - navigating disagreements effectively by acknowledging emotions involved and finding solutions that satisfy all parties.

Overall, Influence in the context of emotional intelligence is about leveraging emotional insights to guide others positively and achieve desired outcomes.

Definition 3.20. ***Communication*** *is defined as the ability to convey thoughts, feelings, and ideas effectively while also being sensitive to the emotional dynamics of interactions.*

Communication is a crucial component of emotional intelligence overall, and Social Regulation in particular. The key components of this sub-construct are graphically represented in Figure 3.21 and include:

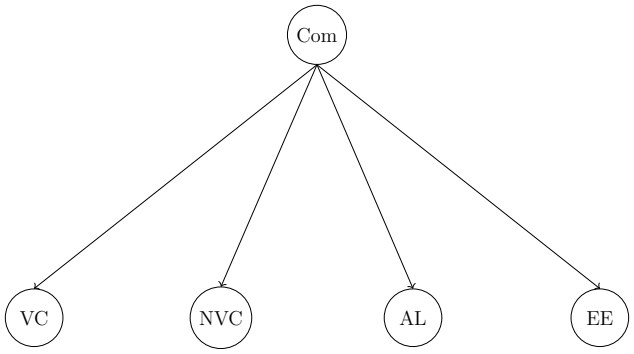


Figure 3.21: TTI Success Insights Communication Sub-construct

1. Verbal Communication - clearly articulating thoughts and feelings in a way that is easy for others to understand, using appropriate language and tone.
2. Non-verbal Communication - understanding and utilizing body language, facial expressions, and gestures to reinforce verbal messages and convey emotions accurately.
3. Active Listening - fully engaging with others by attentively listening, showing empathy, and responding appropriately, which fosters deeper connections and experiences.
4. Emotional Expression - being able to express one's own emotions appropriately, which can help others feel comfortable sharing their feelings and experiences.

Effective Communication within emotional intelligences involves not only transmitting information but also creating a supportive and emotionally aware dialogue that enhances relationships and collaboration.

Definition 3.21. ***Conflict Resolution*** *is defined as the ability to navigate disagreements and tensions in a constructive manner, using emotional awareness and interpersonal skills.*

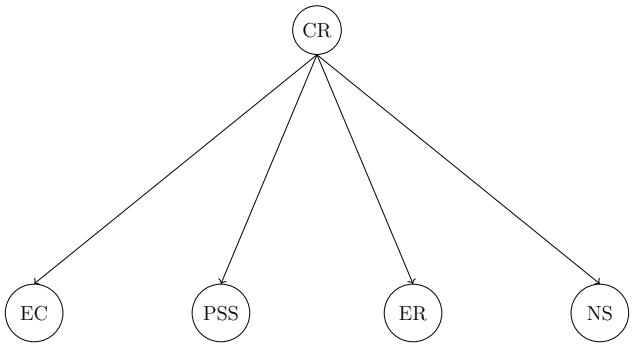


Figure 3.22: TTI Success Insights Conflict Resolution Sub-construct

Conflict Resolution is a critical component for anyone in a leadership role. The key components of this sub-construct are graphically represented in Figure 3.22 and include:

1. Effective Communication - articulating one's thoughts and feelings clearly and respectfully, while also encouraging an open dialogue among all parties.

TTI SUCCESS INSIGHTS®

2. Problem-Solving Skills - collaboratively identifying solutions that address the needs and interests of all parties, rather than focusing solely on positions.

3. Emotional Regulation - managing one's own emotions to remain calm and composed during conflict, which can help prevent escalation and promote a more rational discussion.

4. Negotiating Skills - finding common ground and facilitating compromises that satisfy everyone involved, fostering a sense of collaboration.

Overall, Conflict Resolution in the context of emotional intelligence is about leveraging emotional insights to transform disputes into opportunities for growth, understanding, and stronger relationships.

Definition 3.22. ***Visionary Leadership*** *is defined as the ability to inspire and guide others toward a shared vision while effectively managing emotions within a team or organization.*

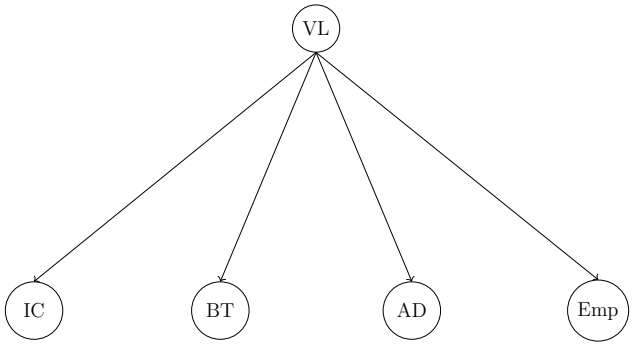


Figure 3.23: TTI Success Insights Visionary Leadership Sub-construct

Visionary Leadership is crucial to leading the teams and organizations of the future. The key components of this sub-construct are graphically represented in Figure 3.23 and include:

1. Inspiring Communication - clearly and passionately communicating the vision in a way that resonates emotionally with others, igniting enthusiasm and commitment.

2. Building Trust - creating a safe and supportive environment where team members feel valued and heard, which encourages buy-in and collaboration toward the vision.

3. Adaptability - adjusting strategies and approaches based on feedback and changing circumstances, demonstrating flexibility while remaining focused on the overarching vision.

4. Empowerment - encouraging others to take initiative and contribute their ideas and talents, fostering a sense of ownership and alignment with the vision.

Overall, Visionary Leadership through emotional intelligence is about creating a compelling and inclusive vision, motivating others to pursue it, and cultivating an environment that supports growth, collaboration, and innovation.

Definition 3.23. ***Catalyzing Change*** *is defined as the ability to facilitate and support transformation within an organization or team while managing emotional dynamics effectively.*

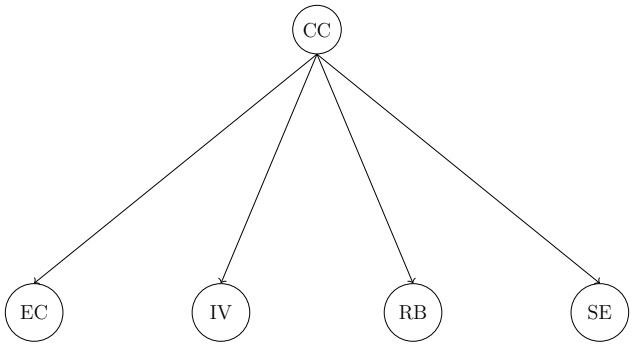


Figure 3.24: TTI Success Insights Catalyzing Change Sub-construct

Being a change catalyst is essential for managing an organization or team through major transition periods. The key components of this sub-construct are graphically represented in Figure 3.24 and include:

1. Effective Communication - clearly articulating the reasons for change, its benefits, and the process involved, while addressing the fears and uncertainties in an open and honest manner.
2. Inspiring Vision - creating a compelling narrative around the change that motivates and excites others, helping them see the positive possibilities.
3. Relationship Building - cultivating strong connections and trust within the team, which can ease the transition and encourage collaboration throughout the change process.
4. Support and Encouragement - providing ongoing support and recognizing individual contributions, which helps sustain momentum and fosters a sense of agency among team members.

Overall, Catalyzing Change through emotional intelligence is about guiding others through transformation with sensitivity and clarity, creating a resilient environment that embraces growth and innovation.

Definition 3.24. ***Building Bonds*** *is defined as the ability to develop and maintain strong, positive relationships with others, characterized by trust, empathy, and collaboration.*

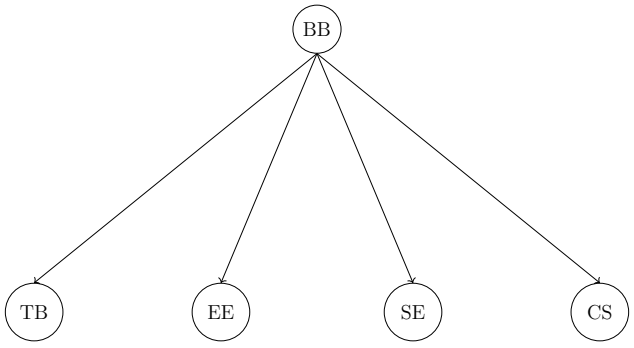


Figure 3.25: TTI Success Insights Building Bonds Sub-construct

TTI SUCCESS INSIGHTS®

The ability to bond with individuals is a critical component of leading individuals and fosters a more impactful ability to influence them as part of a team. The key components of this sub-construct are graphically represented in Figure 3.25 and include:

1. Trust Building - establishing reliability and integrity through consistent actions and open communication, which strengthens relationships over time.
2. Emotional Expression - sharing one's own emotions appropriately, which helps create authenticity in relationships and encourages others to do the same.
3. Support and Encouragement - offering help, recognition, and positive reinforcement, which reinforces bonds and promotes a sense of community and collaboration.
4. Cultural Sensitivity - being aware of and respecting diverse backgrounds and perspectives, which enhances inclusivity and strengthens connections among team members.

Overall, Building Bonds through emotional intelligence is about creating a network of strong, supportive relationships that foster collaboration, trust, and a positive organizational culture.

Definition 3.25. ***Collaboration and Teamwork*** *is defined as the ability to work effectively and harmoniously with others to achieve shared goals.*

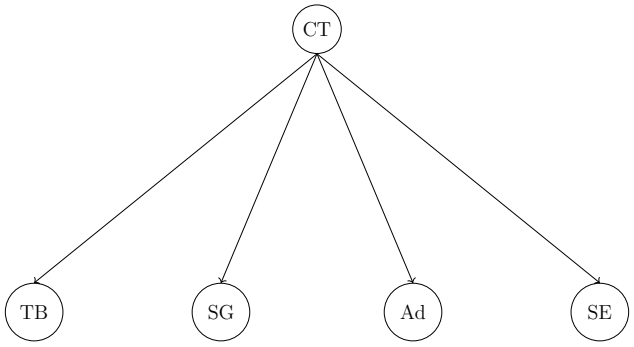


Figure 3.26: TTI Success Insights Collaboration and Teamwork Sub-construct

Being skilled in Collaboration and Teamwork is essential for effective team management. The key components of this sub-construct are graphically represented in Figure 3.26 and include:

1. Trust Building - creating a foundation of reliability and openness among team members, encouraging collaboration and reducing friction.
2. Shared Goals - fostering a sense of common purpose by clearly defining objectives and encouraging participation in decision-making, which enhances commitment.
3. Adaptability - being flexible in roles and approaches, willing to adjust to the needs of the team and embrace diverse contributions.
4. Support and Encouragement - recognizing and valuing the contributions of others, promoting a culture of appreciation and motivation.

Overall, Collaboration and Teamwork through emotional intelligence is about leveraging emotional insights to create a cohesive, effective team environment where individuals can thrive collectively and achieve their goals.

4 A Few Comments on Reliability and Validity

The authors of this manuscript feel that it is important to present some thoughts on the current state of determining acceptable levels of evidence of reliability and validity for our assessments specifically and to psychometric assessments in general. This is a difficult topic to bridge given that we live and breathe in a space that in large part is academic in nature while in practice is a business. Given that it is a business, we tend to look to the academic community for the latest in guidance on such topics as relevant measures of evidence of internal consistency or how to best measure various aspects of evidence of validity.

In large part, we have chosen to follow the American Psychological Association's guidelines set forth in their three-volume series *APA Handbook on Testing and Assessment in Psychology*, see [75], [76], and [77]. We do so not because we are in the business of psychological assessments. Rather, we are in the business of assessments that historically have their roots and early development in fields closely related to psychology. Therefore, it seems a natural fit to follow the guidelines set forth in the aforementioned APA series, as well as other guidance such as that found in *The Standards*, see [4]. Similar guidance is available from such organizations as the European Federation of Psychologists' Associations (EFPA) and the British Psychological Society (BPS). Additionally, there are myriad academic publications from which to obtain information.

Among all the guidance provided by these noted organizations and academic literature, there appears to be a common theme: a lack of true guidance as to what constitutes adequate levels of evidence of reliability and validity. As an example, we provide the following quote. For context, this is from the first volume of the APA series and is related to the chapter on validity, see [163], pp. 67.

> Selecting criteria that are consistent with the theory underlying the construct is easier said than done. Valid external criteria are hard to obtain, expensive, or both. In other cases, valid external criteria may simply not exist. Often, collecting criterion data is impractical, and even when such data are gathered, they may be of questionable reliability or corrupted by biases (e.g., supervisors' ratings used in an employment setting or teachers' ratings in an educational setting).

This quote is a bit disconcerting. In particular, TTI Success Insights has experienced a great deal of demand in the secondary education space in recent years. The main indicator of performance used in that space are students' grades, which are directly assigned by the teachers, and we are being told that such data are likely corrupted. We then pose the following question. Is it possible to validate the use of any assessment in an educational setting? This is not meant to sound alarmist. Rather, it is intended to spark debate on the issue. The TTI Success Insights suite of assessments has been used with a great deal of success in this arena, yet we are faced with a seemingly unsurmountable obstacle. We will come back to this part of the discussion later.

The remainder of this manuscript presents some of the obstacles we see coming out of the academic world in the various areas of reliability and validity. We view the relationship with the academic community as a very important one. The academicians develop the theories that assessment developers and users apply in their everyday work. We seek to strengthen that relationship rather than create a divide.

4.1 Thoughts on Evidence of Reliability

We go more in-depth into the history of reliability coefficients in the section on internal consistency in [28]. For purposes of this section, we mention a few basic historical facts and leave the rest for another time. The concept of reliability of an assessment has been around since at least 1937, when Kuder and Richardson introduced the so-called α coefficient for dichotomous items, see [105]. It is not the intent of the previous statement to give a complete treatment of the history of reliability, only to establish that the concept is not a new one.

The main issue we wish to discuss here is the lack of a consensus in the world of assessments as to what is the best approach to measure reliability and at what levels may we consider an assessment reliable, or rather, to have an acceptable amount of evidence of reliability. There have been many published articles in peer reviewed journals discussing the differences of opinions on this topic. Two of the more recent ones are [44] and [96]. [44] is a bit more useful in that it discusses several different approaches and offers some suggestions as to directions one could take. [96] is a nice exposition on an empirical analysis of the use, or misuse and misunderstanding, of the α coefficient specifically.

[44] offers opinions based on six common misperceptions surrounding the α coefficient and goes on to suggest ways of using structural equation model based techniques (SEM), such as variants of McDonald's ω, to compute reliability estimates. Their six misperceptions are:

1. α was first developed by Cronbach;
2. α equals reliability;
3. A high value of α is an indication of internal consistency;
4. Reliability will always be improved by deleting items using "alpha if item deleted" analysis;
5. α should be greater than or equal to 0.70 (or, alternatively, 0.80);
6. α is the best choice among all published reliability coefficients.

In a similar fashion, [96] presents six alleged misunderstandings and then conducts an empirical analysis of published authors to determine whether the alleged misunderstandings are accurate.

1. α is equal to the reliability of a test score;
2. The value of α is independent of the number of items of a test;
3. α is an indication of the unidimensionality of a test score;
4. α is the best choice among reliability coefficients;
5. There is a particular level of α that is desired or adequate;
6. If removing an item increases α, the test is better without that item.

We are not here to pass judgment on the correctness of the arguments nor the adequacy of the empirical study mentioned above. The reader interested in the arguments or the final analysis may consult the original works in [44] and [96], respectively. We do, however, want to point to a small handful of comments made by the authors of both works that are particularly revealing about the current state of affairs when it comes to internal consistency measurement.

The first issue we wish to bring into the discussion is the idea that Cronbach is the originator of the so-called α coefficient. Though [44] and [96] present slightly differing views, they generally agree that [105] first proposed a "reliability formula (called KR-20) that can be used for data on dichotomously scored items...", see [44]. [96] states that "Alpha was introduced by Kuder and Richardson (1937) for dichotomously scored items". We point out this difference given that [44] does not refer the result of [105] as "alpha" while [96] does. This is not a major issue in our opinion, but does point to an inconsistency in definitions in the literature that may need to be addressed. Are the formulas for KR-20 and alpha the same, or at least can they be considered close enough so that this inconsistency is not worth discussing?

It is our understanding that the two measures are equivalent *if they are applied to dichotomous data*. However, this does not imply they are equivalent. In fact, KR-20 should not be applied when data are not dichotomous. In our opinion, this difference is important and should be highlighted, especially in light of the fact that the publication of origination of this discussion is one highlighting inconsistencies in the understanding and use of the α coefficient.

[44] traces the history of reliability coefficients to Spearman and Brown, both independently publishing in the same issue of the *British Journal of Psychology*, see [165] and [36]. The formula published has since become known as the Spearman-Brown formula. [96] begins their discussion with [105]. We would like to point out that our opinion is that both articles are informative and that we are using this likely minor inconsistency to further and extend the debate that [44] is introducing.

Table 4.1: Common Reliability Coefficient Interpretation

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

The debate is centered on what is the best measure of internal consistency. The debate is clearly not settled given that [44] was published in 2015 and [96] was published in 2019. Our attempt to further the debate is an attempt to get clarification on what is the best approach to establishing acceptable levels of internal consistency. Our extension is to point out the lack of guidance in all areas of evidence of reliability and validity.

As another example of issues we see, consider the discussion on acceptable levels of a measure of internal consistency. For example, the Table 4.1, above, is presented in [63] and reproduced here for purposes of the current discussion. Additionally we present the following quote from [12]:

> How high should values of α (or related indices such as KR-20) be? I hesitate to give an answer to this question because estimates of α can be affected by many things, such as the purpose of the test, the heterogeneity of the sample from which it was obtained, the conditions of testing, and the number of items.

Further, we have [130] stating that 0.70 is acceptable during development, 0.80 for "basic research scales", and 0.90 for clinical settings are appropriate minimums. [144] recommends a minimum of 0.80 without the same restrictions as [130]. As a final example, [101] suggests values of 0.95 for clinical settings. [12] concludes her remarks on recommended levels of reliability with:

> The higher values for clinical decisions reflect the serious consequences of such decisions and the need for correspondingly higher standards. The same argument for higher reliabilities could be made in the context of educational tests that are used for student placement, graduation, or remediation decisions or for licensure or certification tests. Indeed, any test that will be used to make consequential decisions about students, clients, patients, or others, should contain as little error as possible.

We agree with the idea that our measurements should contain as little error as possible. We agree that it is in the best interest of all involved (the developer, the user, the respondent, etc.) for any assessment to have as little error as possible. The issue we wish to discuss is whether it is possible to pose a well-defined response to the concept of as little error as possible while, at the same time, not bringing our industry to a standstill waiting for a solution.

We have attempted to show the lack of agreement in to major areas of measuring internal consistency. First, there is no agreement across the academic community as to the best measure of internal consistency. Second, there is no agreement on what levels constitute acceptable levels in the process of establishing evidence of reliability. If this is the case, what is the motivation for the practitioner or user of such instruments to extend their own knowledge in an attempt to establish such evidence of internal consistency? If those largely responsible for establishing the standards are either unable or unwilling to state, definitively, what those standards are, how can the same groups stand so definitively for what they are not?

4.2 Thoughts on Evidence of Validity

We see similar issues in the arena of the various evidence of validity requirements. Depending on whom one chooses to follow, one may be dealing with differences in constructs (both number and type). The good news is that while there are some differences, there is usually a natural mapping between them. The not so good news is that we still appear to have a lack of consensus as to what constitutes acceptable levels of evidence of validity, whatever that may mean, i.e., depending on the context.

In some sense, that last comment sums up the issues we see with some of the ideas in both the reliability and validity spaces. Perhaps it is the choice of words used to describe the concept. Perhaps it is that the concepts themselves are not as well-defined as some assume them to be. Well-defined seems to be a discussion topic for a different time, but we felt strongly enough about the idea to at least mention it. However, the choice to use the words reliability and validity does appear to pose a serious problem.

Perhaps an example from a previous life of one of the authors is relevant to the discussion. At one point in time one of us spent several years working in mathematical and statistical modeling largely in the banking and finance industry. Various modeling approaches in that industry have been around for a long time and various efforts at global regulation go back at least to the first Basel Accord (Basel I) in 1988. We note that the Basel Committee on Bank Supervision (BCBS), which commissioned Basel I, II, and III, was formed in 1974, implying the discussion has been around for much longer. As a side note, Basel, Switzerland is the location the BCBS was founded in and calls home.

These accords are not regulatorily binding in the U.S. However, the Federal Reserve Bank and Office of the Comptroller of Currency in the United States routinely adopt the models outlined in the Basel Accords as part of their monitoring and supervision approaches. This leads to the advent of a regulatory approach called model validation. Model validation is really intended to be a peer review process focused on helping to improve an institution's mathematical and statistical modeling approaches. However, in the experience of one of these authors, it often led to confrontational relationships between the model validation teams and the model building teams and/or model owners. This was especially true when the model validation team came from an outside source such as an external consultant or government auditor.

This led one of these authors to suggest to the company he worked for a change in use of language from a model validation project to a peer review project. That simple change in language made huge differences in the relationships and proved highly productive during that time. While it is understood that such a change in language in this space is a really big ask, perhaps it is worth the discussion.

We do not believe that the intent of the words reliable and valid was to have a black and white cutoff point, nor do we believe it was intended to be part of answering the question "Is your assessment reliable and valid?" We choose to follow the APA guidance as presented in [75] and use the language evidence of reliability and validity. However, as a company we receive inquiries on an almost daily basis asking if our assessments are reliable and valid.

We do note that [163] differentiates between validity and validation, noting that "the former refers to the degree to which an assessment fulfills its intended purpose" while "validation refers to the process of gathering and reporting evidence to evaluate the use of a test for a particular purpose". While this distinction is noted here, we doubt the average user, potential client, or test taker has the time to take to understand the difference.

[4] states that there are five sources of evidence of validity. However, we note that *these are not defined as* ***the*** *sources of validity*, see pp. 13:

> The following sections outline various sources of evidence that *might be used* in evaluating the validity of a proposed interpretation of test scores for a particular use.

The italics in the previous quote are added by the authors of this manuscript for emphasis. We make note of two important implications of the previous quote. The first is the use of the phrase "might be used" implying that there are potentially other sources of validity. The second is that the use of the phrase "evaluating the validity of a proposed interpretation of test scores..." This second is a statement about the nature of the relationship between validity and an assessment. It is not the assessment itself that is valid, rather it is the interpretation and use of the scores that is of importance.

The five sources are listed as follows:

1. Evidence based on test content;
2. Evidence based on response processes;
3. Evidence based on internal structure;
4. Evidence based on relations to other variables;
5. Evidence for validity and consequences of testing.

Each of these areas is intended to measure a specific aspect of the use and interpretation of assessment scores. An example provided in [4] for test content evidence is as follows, see pp. 14-15. They state that if we are measuring mathematical ability at a certain level, it is fair to test on all mathematical content up to and including that level whether the concepts have been introduced to the respondents or not. In a different scenario, suppose we are interested in testing the knowledge base of that same group of students based on delivered curriculum. In this case, the content domain should be limited to what the students have been exposed to in the delivered curriculum.

We are not dealing with mathematics curriculum. We are dealing with, for example, Style Insights® and four constructs. We are interested in the content of the test (the questions, items, frames, etc.) and the construct(s) it is intended to measure, see [4], pp. 14. We therefore need to have a serious conversation about what the construct definitions are (for all our constructs on all our assessments) and then make a determination as to whether the content domain of definition is adequately covered without introducing irrelevant information. These two concepts are known as construct representation and construct irrelevance.

Conceptually speaking, evidence of validity based on test content is relatively easy to understand. In practice, it may be quite difficult to establish. As an example with Style Insights, the history going back to Marston has been well documented by many. However, there do exist some issues that prevent a full content validity evidence claim based on this lineage. The arguments laid out in Marston's work have been debunked by both the psychology and neurology communities. The understanding of how the brain facilitated emotions in the early part of the 20^{th} century has been shown to be incorrect by modern technological advances such as the electroencephalogram (EEG).

Further, most of the development of the current iteration of the DISC assessment occurred in the late 1960s to early 1970s by Geier who never published his findings. These two facts place a difficult obstacle in the path of any content validity argument. This is not to say that one cannot create an argument for the content validity of the DISC assessment. It means that we need to find or develop the theory of a four-factor model that is not based directly on the work of Marston and Geier.

Some may challenge our assertions here. Simply put, the highest form of validation comes from an independent third party review of our assessments, their use, interpretations, and so on. This third party reviewer is likely to be some form of academician, and we must approach a project of this type with the utmost transparency if we expect to be treated fairly. We need to own the negative and establish processes to correct that which is correctable or find different approaches and justification for that which is not directly correctable.

Evidence based on response processing is a challenging topic. The APA takes the following position on this portion of validity evidence, see pp. 76 of [163].

> Gathering validity evidence based on response processes is perhaps the most difficult validity evidence to gather because it involves demonstrating that examinees are invoking the hypothesized constructs the test is designed to measure in responding to test items. ... Gathering evidence is difficult because one cannot directly observe the cognitive processes going on within people's heads as they respond to test items.

Or, can we? In some sense, there seems to be a pattern in which we are asked to gather certain kinds of evidence of validity followed by commentary on how difficult, if not impossible, gathering such evidence is. For example, see the introduction to this chapter and its discussion about criterion data. We do agree that gathering this type of evidence is challenging, and it is doubtful that many, if any, assessment developers (academic or otherwise) are even attempting to do so.

However, TTI Success Insights has had an active neurological research group since 2011, and one of the areas of recent focus has precisely been studying respondent brain reactions to our assessment items. The current study has focused on the TTI Success Insights Emotional Quotient assessment with an anticipated peer reviewed journal submission sometime in late 2020 or early 2021. We have already had the protocol paper accepted and published, see [29].

The APA offers the following suggestions in light of their statement on the difficulty of gathering evidence based on response processing.

1. Think aloud protocols and cognitive interviews;
2. Chronometric analysis;
3. Evidence-centered test design;
4. Mathematical modeling of item difficulty;
5. Evaluating processes used by graders;
6. Other evidence based on response processes.

Several of the listed approaches are non-starters for TTI Success Insights. For example, think aloud protocols may be an interesting approach, but we could do no more of that type of evidence gathering than we could with gathering brain images during item responses. Perhaps an interesting study could come of this, but it would be quite limited.

At the current time, TTI Success Insights does not have the ability to record time to respond to individual items on our assessments. We do record total time, but that is not what chronometric analysis requires. However, TTI Success Insights is currently working on many projects to update our database and computerized assessment approaches. One such project will bring the ability to record individual item response times. When that time comes we will be in a position to perform such studies.

For evidence-centered design, this approach is, by definition, part of the design process of an assessment development. If we ever completely design a new assessment from the very beginning, we should incorporate this type process. Otherwise, it is not of use to us at this time. Similarly, evaluating processes used by graders is meaningless in our setting as we do not grade anything.

We are then left with the Other category and mathematical modeling of item difficulty. Our brain research falls into the Other category and is similar to some suggestions given in [163] such as monitoring eye movements during task performance. As for mathematical modeling of item difficulty, TTI Success Insights has incorporated GRM approaches from IRT in our Likert style response assessments (e.g., Emotional Quotient). We have not specifically viewed these models from the assessment development standpoint and "treated item attributes as facets", see [163] pp. 77-78. We are using these models for two purposes. First is to generate a weighted scoring model for each scale measured by the assessment and second is to gain diagnostic information on the assessment scales.

The extension of these type models to our forced-rank assessments is more challenging. Our plan is twofold. We are first gathering data on the DISC and Motivators assessment items in Likert style formats. We plan to use graded response models or similar to model item difficulty and discrimination and use that information to determine optimal, in some sense, pairing of D, I, S, and C items. See the section on graded response models applied to Likert response data from our

DISC assessment for a more complete explanation, [28]. Once these pairings have been established, we will build forced-rank frames and gather response data on these new frames. The second IRT approach will take the (possibly) updated frames and follow a Thurstonian IRT model, or similar, to analyze the data.

A more complete presentation of the Thurstonian IRT model, including references, is provided in the section on Thurstonian IRT models applied to forced-rank assessments in [28]. Briefly, one cannot directly apply IRT models to forced-rank assessment data. Therefore, we transform the data using the Thurstone's Law of Comparative Judgment. This new data may be used to estimate the parameters of a multidimensional IRT model. At this time, our approach has been to use a multidimensional probit model. A probit model is another name for a model based on a normal or gaussian distribution.

Validity evidence based on internal structure appears to be the area of validity for which the most tools are available and also appears to be the the area most readily implemented. Tools such as exploratory and confirmatory factor analysis are available and described beginning on pp. 73 of [163]. Also discussed are IRT models, multidimensional scaling, and evaluating the invariance of the assessment structure through the use of differential item functioning (DIF). This last area is an important one in the US for legal reasons. DIF and similar tools are intended to measure the invariance of the response patterns of different groups of individuals. For example, the concept of disparate impact in hiring processes may be partially addressed by showing that protected classes, such as gender, age, race, etc., are not negatively impacted by assessment scores.

That is not to say that there aren't differences in different groups of individuals. Differential item functioning analysis is intended to identify where the differences occur internal to the assessment items. It is then up to the assessment developer to provide guidance on how to compensate for such differences, if they exist, during the use of the assessment scores.

All the aforementioned approaches are implemented for the various TTI Success Insights assessments as they apply. As an example, like IRT models, factor analytic models are not directly applicable to forced-rank data. The reason for this is technical, and we briefly mention it here. Factor analysis is also known as analysis of covariance. These processes essentially compare a theoretical covariance matrix, with certain assumed structure, with the population covariance matrix of the assessment data. During the estimation process, it is necessary to invert the population covariance matrix. Unfortunately, it can be shown that any forced-rank assessment generates a singular covariance matrix. In simple English, one cannot invert a singular matrix. It is essentially dividing by zero.

As mentioned previously, we plan to implement the Thrustonian IRT model to help in the analysis of both the response processing and the internal structure evidence of validity. We do note that, for diagnostic purposes, we have employed exploratory factor analytic approaches on the Style Insights scales individually. While this approach is not directly applicable to establishing internal structure validity, it has been invaluable in identifying those items that perform well and those that do not. What it cannot do is give us a reason for why certain items underperform given that they are also influenced by the existence of three other items in the forced-rank frame. See Section 4.5 for a discussion of the pros and cons of both forced-rank and Likert style scales.

Before going on to the next area of evidence of validity, it is important to note that while factor analysis is a very commonly employed tool in assessment development and there is a host of published articles and books on the topic, there is little guidance on what is a "good" factor analysis. So we factor analyze our assessment data. Now what? The only guidance we have found on this

topic is to look at a couple basic things. We would like to have simple structure in the factor pattern matrix, and we should look at the total variance explained by scale. What is simple structure of a pattern matrix? What is an acceptable level of total variance explained? The following is guidance from [4], Standard 1.13, beginning on pp. 26.

> It might be claimed, for example, that a test is essentially unidimensional. Such a claim could be supported by a multivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the score variability attributable to any other identified dimension, or showing that a single factor adequately accounts for the covariation among test items.

We can infer from the literature essentially what we desire in the form of a factor pattern matrix. However, what we cannot infer is what is an acceptable level of factor loading or what is an acceptable level of variance explained. To date, these authors have found no definitive statements on either of these topics. It would seem reasonable that 50% of the variance could be used as a minimum. It appears, however, that no one is willing to definitively state that this is a reasonable *minimum* level of variance explained to aim for. We emphasize the word *minimum* to clearly indicate that should be a minimum requirement and not a final goal.

We do note that part of the issue in making definitive statements on a topic such as acceptable levels of factor loading is that it varies with the number of items associated with the construct being measured. The variance explained computation based on a small number of assessment items cannot absorb a weakly loaded item well while the same computation on a large number of items can. As mentioned in the section on reliability above, more definitive guidance from the academic community on topics such as this would be appreciated.

We briefly discussed the evidence of validity based on relations to other variables in the introduction to this section. In Section 4.3, we discuss correlation versus causation and take a more in-depth look at some of the issues we see with this approach. We briefly summarize a key point. The social sciences community appears to have a much lower standard for what are considered acceptable levels of correlation between variables than, say, the physics or mathematics communities. This seems to cause a problem. Not only are the discussions in the social sciences using correlation as a proxy for causation, which is dangerous in its own right, but these discussions also accept quite low levels correlation as explanatory.

Finally we discuss evidence based on consequences of testing. This evidence is based on an evaluation of the intended and unintended consequences associated with a testing program. We provide the following quote from [163] as evidence of the lack of consensus in the community for this type of evidence.

> Whether validity evidence based on consequences of testing is relevant in evaluating the validity of inferences derived from test scores is a subject of some controversy.

To be fair to the authors of [163], they do take the stance that

> We believe this debate to be one of nomenclature, and given that virtually all testing programs have consequences on some level, it is important to evaluate the degree to which the positive outcomes of the test outweigh any negative consequences.

Our point here is that while the APA's viewpoint is that evidence based on consequences of testing is an important measure to be considered in the overall validity argument, the APA does not review assessments in a manner similar to the BPS. Given that everyone is not in agreement, how much time and effort does one expend on an area in which there does not appear to be a consensus?

4.3 Correlation v. Causation

One would be hard pressed to find anyone who has not heard the phrase "correlation does not imply causation". As a simple example of what is meant by this catch phrase, consider Figure 4.1. The correlation between the two variables under consideration is just under 0.95 (0.9471 to be exact), see https://www.tylervigen.com/spurious-correlations for the original plot. In this case, we are comparing the relationship between per capita (U.S.) cheese consumption and the paucity of individuals meeting an untimely death due to inadvertently becoming strangled in their own bedsheets.

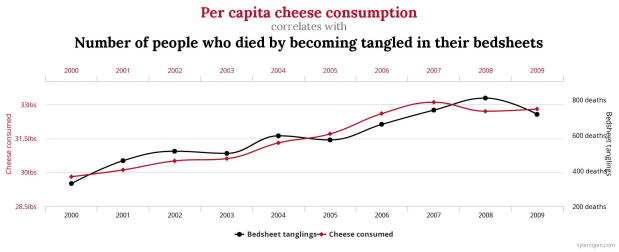


Figure 4.1: Bad Correlation v Causation Example

This is a clearly ridiculous example, yet correlation, and not necessarily causation, between variables is one of the most common tools used to establish evidence of validity based on relationships to other variables. It seems odd to accept correlation as evidence of a causal relationship while claiming that correlation cannot establish the desired causation. We noted in the introduction of this section the APA's comments on criterion related data. In this section we discuss the apparent contradiction that is the use of quite low levels of correlation to help establish a causal relationship between assessment scores and external variables.

The authors of this paper are relative newcomers to the world of social sciences. We have each spent most of our professional lives working in the so-called hard sciences. As noted earlier in this manuscript, one of us spent many years working in mathematical and statistical modeling and model validation in the banking and finance sectors. One general rule of thumb or guideline is that, in practice, correlation less than $|0.40|$ between variables essentially means there is no relationship between those variables.

Our viewpoint is that when attempting to build a predictive model, we would only keep explanatory variables that were reasonably well correlated with the response variable and were correlated at less than $|0.40|$ with the other potential explanatory variables. The thought here is that correlation really only begins to become meaningful when greater than $|0.40|$. Another way to think of this is if we square the correlation coefficient, we get an approximation of the linear variation accounted for between the two variables. For a correlation at the level of 0.40, only 16% of the linear variation between the two variables would be accounted for.

It is not the intent of this section nor is it the intent of TTI Success Insights to point to specific competitors or individuals on this topic. However, a review of several competitors' claims of evidence of validity based on relationships to other variables shows that in many cases, claims of the existence of such relationships is established based on correlation coefficients anywhere from 0.10 up to 0.50. The majority of those reviewed by the authors of this work are well under 0.30.

Putting this into the context of explaining the linear variation between the variables, we see that claims are made for the existence of (causal) relationships between variables based on 1% up to

9% of said linear variation being explained. It is the hope of TTI Success Insights to help establish that explaining less than 10% of the variation between variables establishes a fairly weak argument for correlation and really does not establish anything in the area of causality.

We do understand that traditionally lower levels of correlation have been accepted in this area, and we make no claim that all work should be thrown away because of this acceptance. We do argue that as the future unfolds, we look to raise the standards of what is considered acceptable levels of correlation to help establish causal arguments. We find it odd that the acceptable level of variance explained for a scale on an assessment is at a minimum of 50% (at least by some suggestions), while in the same field less than 10% may be thought of as providing an acceptable level of information. This is clearly not a complete apples to apples comparison. However, the interpretation should be clear from the context.

In the spirit of not pointing out a problem and then offering no solution, we offer at least a simple proposal for a possible better approach to establishing relationships between variables. Rather than using correlation coefficients, we can use other modeling techniques that may be scored in ways not directly related to correlation. For example, one of the approaches used at TTI Success Insights is to establish relationships between set inclusion based on an external variable and the odds of existing in certain scoring levels on our assessment scales. We do this using a logit model, although many other possibilities exist. We then measure the success or failure of the models based on a contingency table style analysis including the use of the receiver operating characteristic (ROC) curve. We note this is not the only approach, just one that has proven successful for TTI Success Insights.

Essentially we establish that our variables provide a better than random chance of identifying randomly selected groups of individuals from a desired target population out of a larger population. This analysis would be more enlightening were it to be combined with some form of performance or longevity metrics, at a minimum. In those cases where we have performance data, we have seen improved performance of our modeling approach in identifying not only the individuals, but also accurately ranking these individuals based on their performance.

This is not a perfect solution to the aforementioned problem. However, it does extend the analysis and moves away from using a tool that is not a good fit for the conclusions being drawn.

4.4 Understanding Self-report Assessment Bias

In this section we discuss some of the published strengths and weaknesses of the self-report format of assessments. Instruments that rely on self-report require the participant to rate themselves by answering open-ended questions or by indicating whether or how well a statement describes them. Primary self-report methods are surveys and interviews. Key strengths of self-report methods are that they can be relatively easy to administer and score. These tools can tap directly into the test taker's personal experiences, thus gaining insights into such areas as motivation, communication styles, behavioral preferences, self-perception, ideal work environment, and a host of other personal attributes. Even the medical field relies on self-reporting as a first step during diagnoses.

It is important to note that these different assessment purposes/applications require different criteria for judging validity and reliability. Some purposes are relatively low stakes or formative in nature, e.g., demographic data, enhancing self-awareness, or measuring for professional development. Other purposes are higher stakes, such as part of hiring, promotion, or retention decisions. The following five work-related purposes for measuring personal attributes may be found in [161]:

1. Formative feedback and career guidance
2. Career assessment
3. Program evaluation and practice improvement
4. Personnel selection and evaluation
5. Readiness certification

Each of these assessment applications bring with it diverse consequences resulting in different ethical and legal concerns. Therefore, it is vital that the assessment users understand both the strengths and limitations of self-report assessments.

The key to solving any problem is to first clearly identify the concerns. [16], [115], and [136] have all described the psychometric properties of existing self-report measures that need to be acknowledged and, if possible, addressed during any continual improvement process, as well as when interpreting and presenting results. While identification of concerns is crucial, it is important for the reader to understand that TTI Success Insights is not only aware of such concerns, but is also continually striving to rectify these concerns whenever possible. A complete discussion of how each potential bias is being addressed by TTI Success Insights is beyond the scope of this manuscript, a follow on manuscript addressing each of the following areas is in progress.

As outlined in [16], several particular biases that can influence self-report measures include:

1. Social desirability bias - the tendency to respond in ways that, rather than truthful, are culturally or socially appropriate, acceptable, or desirable.
2. Reference bias - the tendency to evaluate oneself in comparison to one's own peer group rather than to a broader or more objective set of standards. This most likely occurs due to an individual's lack of knowledge about groups beyond their own peers.
3. Acquiescence bias - the tendency of survey respondents to answer all questions on a survey in the affirmative.
4. Consistency motif bias - stems from the tendency of survey respondents to want to maintain what they consider to be consistency in their pattern of responses. Respondents may apply their own theory of how test items are interpreted and adjust their responses accordingly.

Faking may also be a problem in self-report methods, especially when the test taker knows that the stakes are high. When using self-report methods, test users may mitigate biases and faking through appropriate strategies of test constructing or by triangulating sources of evidence.

A review of [5], [11], [68], and [177] provide additional concerns for which we are aware and prepared to rectify whenever possible. While most of the following are incorporated to some degree in the core challenges listed above, it is useful to examine any issue from multiple perspectives. The following is a partial list of some of the alternatives described in the aforementioned literature.

1. Introspective ability - describes the lack of introspective ability to respond completely accurately to assessment questions. This bias is a concern when examining constructs such as self-awareness. It may not be appropriate to assume that everyone has an adequate level of awareness to be able to accurately respond.

2. Honest/Image management - self-reports rely on the honesty of the respondent and the research must trust the responses to be honestly provided.

3. Understanding - respondents may vary in their understanding or interpretation of particular questions.

4. Rating scales - the use of rating scales implicitly assumes that all participants interpret the scale in the same manner. One major issue is that different respondents may rank items differently based on differing interpretation of the magnitude of the point scales.

5. Response bias - the amount of evidence required to affirm or disagree with a statement may vary depending on respondent experiences or any other of a number of variables.

6. Ordinal measures - almost all self-report measures produce ordinal data. Ordinal data only tells the researcher the ordering of the ranking, not the distance between. The concept of ordinal data may be contrasted with interval data which does provide distance information (but still suffers from similar interpretation issues as that mentioned under rating scales above).

7. Control of sample - Assessments that are taken online are especially susceptible to the aforementioned biases. There is virtually no control over the environment of the respondent, time frame, state of mind of the respondent, or any number of other adverse conditions.

4.5 Forced-rank v. Likert-style Assessment Formats

Very loosely speaking, there are two basic response options available in the setting of self-report psychometric assessments. These are the forced-rank and Likert-style response formats. Each of these approaches has its pros and cons, and the authors of this work are not aware of any general consensus on the question of which is the preferred option. In the remainder of this section, we attempt to present some of the various arguments for and against each option, as well as attempting to refute those arguments, especially in the case of the arguments against.

One main argument against the use of forced-rank, and in favor of Likert, is that only intrapersonal comparisons are valid in the setting of forced-rank generated scores. In other words, we really should not make interpersonal comparisons. The argument for why this is the case is based on the fact that we cannot measure how much an individual prefers a given item in a forced-rank frame. It is not possible to tell if there is a unit measurement between each of the four items that one has ranked from 1 (e.g., most like me) to 4 (e.g., least like me).

In the spirit of collegiality, we agree with this argument, to a point. We agree that magnitude of difference for an individual is a short coming of the forced-rank response format. We do not agree that it is not possible to determine the distances in question, and hence interpersonal comparisons are not valid. One possible approach to aid in determining the distances in question and facilitating valid interpersonal comparisons is the Thurstonian Item Response Theory approach mentioned in Section 4.2 and presented in more detail in section on Thurstonian IRT models in [28].

In a nutshell, the Thurstonian IRT process unfolds the forced-rank comparisons and places them in the setting of the Thurstonian Law of Comparative Judgment. In that space, item response theory tools are available for use and an IRT model may be computed. From that stage it is a matter of interpretation of the results and generation of an adequate scoring model, which is actually a natural by-product of computing the model coefficients. Once complete, interpersonal comparisons are completely valid.

In the same argument, it is at least implied that the Likert-style format directly allows for interpersonal comparisons. We partially agree, but also point out that the Likert format assumes that the distance between every choice is automatically the same. In other words, the Likert format assumes that the choice to rank an item 4 or 5 or 6 on a scale (from 1 to 6, say) are all weighted the same. Stated in a more concrete way, the choice to go between Slightly Agree (4) to Agree (5) is weighted the same as the willingness to go between Agree (5) and Strongly Agree (6). In the language of the previous section, Likert-style scales also generate ordinal data.

To propose an even starker contrast, suppose we consider that the same format also weights the choice to move from Strongly Disagree (1) to Disagree (2) the same as the willingness to go from Agree (5) to Strongly Agree (6). The authors of this manual have not seen a strong enough argument that this is, in fact, the case. All questions are not weighted equally across their response possibilities, hence the reason for the existence of such IRT models as the Graded Response Model. In other words, we agree that the Likert-style format is relevant and useful if it is used appropriately and has an adequate scoring model generated with a relevant IRT model or a similar approach that allows for an analytical weighting scheme to be applied.

One argument against the use of Likert-style response formats is that respondents can be all things at all times. This is, of course, true. It appears to fall under the headings of Social Desirability and Acquiescence as noted in the previous section. The so-called "faking" process is also a potential pitfall to be concerned with. We do note that this highlights one of the advantages of the forced-rank format. That is to say that the forced-rank approach does not allow one to be all things. It is also the case that, unless the respondent is extremely familiar with the assessment and its constructs or scales, it is more difficult for the respondent to simply fake it to obtain a particular score profile for a given assessment.

Further, for both the Likert and forced-rank formats, modern technology presents the challenge of the "how do I game the assessment" websites that have popped up over the years. This is particularly problematic if the assessment(s) in question are part of a hiring process or if the results are used for training or advancement at a particular company. These prospects place a great deal of pressure and responsibility on the assessment developer to ensure the development process constructs assessments and/or scoring procedures designed to identify and deal with such behavior. As noted, this is an issue for both response formats.

One longstanding critique of the forced-choice format is that the usual classical test theory (CTT) analytic techniques do not apply to the data generated by such assessments. Further, the extension of classical test theory to item response theory is also not directly applicable to the output of forced-rank assessment response data. It is impossible for the authors of this work to disagree with these comments. In particular, there are mathematical issues with the application of either CTT or IRT to forced-rank assessment data.

The arguments for why this is the case are beyond the scope of this work. However, there are two main points. The first is related to both IRT and factor analysis from CTT. In both cases, analysis requires the use and inversion of the covariance matrix underlying the data, and this is not possible in the case of forced-rank data. Any full covariance matrix based on forced-rank data is not invertible, by definition. The second argument is that even in the cases where one can mathematically apply CTT or IRT approaches, the interpretation of the results is suspect. The example to keep in mind is any forced-rank assessment in which we have n scales. If we take the data generated by any $n-1$ scales, we may have a perfectly valid and invertible covariance matrix. The data in those is still impacted by the removed scale and any interpretation based on the $n-1$

scales needs more justification, if possible.

In a paper published in the *Journal of Occupational Psychology* (BPS) from 1988, see [99], the authors note

> Only by deleting a variable or variables can strict dependencies be removed but it will be appreciated that the variables left still have shared specific variance, etc., and so problems of interpretation remain.

The authors in [99] are generally quite negative towards forced-rank data, but mostly due to the the use of the data generated to justify the reliability and validity of those assessments. In the abstract to [99] the authors at least state the following:

> This is not to say that ipsative tests have no utility but that the claims made for their validity and reliability and their applicability to inter-individual comparisons are misleading.

The authors of this work would like to take this moment to note that this is one of the main themes we are trying to address. The academic community has, over the years, been very vocal about their critiques of assessments without offering solutions to the problems posed. It is not logical to argue with the claims in [99]. It is logical to ask for those who actively research these areas to help pursue solutions to the problems they uncover, not just complain that others do not follow a not well-defined set of rules.

The earlier reference to the Thurstonian IRT process is one approach to addressing this issue as it relates specifically to forced-rank assessment formats. It took some 20 years, but a group of academics took the time to study and uncover possible solutions to the problems pointed out in the 1988 paper [99]. The work presented in [33], [34], [35], [116], and [117], among many others, presents a bridge between the CTT/IRT world and that of forced-rank assessments. The good news is that we see some movement in the academic community to address an issue in that will be to the benefit of the business community, and it is appreciated. The not as good news is that we are still left to interpret the output, meaning the use of these models is still subject to the question of whether the resulting output is *good enough*.

As an example, one possible output of the work previously mentioned is a 2-dimensional surface plot of a so-called item information surface. In theory, this should be analogous to the item information function in the 1-dimensional case. However, there is little even in the 1-dimensional case to provide guidance on what is a "good" item information function. Again, in theory, we should be able to interpret something about the overall reliability of scales in relation to each other, yet it is unclear what is an acceptable level of information.

At TTI Success Insights, we employ both approaches, depending on the intended purpose of the assessment. As an example, in the case of the Style Insights assessment, we are attempting to measure observable behavior in a self-report form. The DISC model is the underlying model of the assessment and has as a basis for behavior the four variables D, I, S, and C. This basis is used to approximate a lower dimensional model to the true behavior model. While the various behavior variables are independent of each other, they may, in some cases, be equally attractive to the assessment respondent. In an attempt to differentiate, TTI Success Insights believes a forced-rank model is best suited to capture such differences.

In contrast, consider the Emotional Quotient model in which the scales are not as independent. What we mean by not as independent is that, in some cases, a scale may depend to a certain extent

on another scale measured by the assessment. For example, it may be the case that one cannot really be able to self-regulate if one does not first have at least a moderate level of self awareness. In this case, it is perfectly acceptable to have these scores correlate at a higher level (as compared to DISC); and, thus, TTI Success Insights believes a Likert-style response format to be the better choice in this case.

4.6 Parting Thoughts

We began the discussion in this manuscript referring to some comments related to relationships to other variables evidence of validity made in the APA series on testing and assessment in psychology. Our intent was to not only spark some debate, but also point to some apparent contradictory approaches taken in the assessment world in general.

The comments related to criterion data are centered on two areas. The first is grades in an educational setting. The second is related to supervisor ratings in an employment setting. We would like to finish this manuscript with a brief discussion of the first.

For many people (in the U.S.), assessments are part of every day life, at least hyperbolically speaking. In the U.S., most of secondary education is focused on passing tests of one kind or another. Students are faced with everything from assessments that grew out of the so-called No Child Left Behind movement of the early part of the beginning of the 21^{st} Century in the U.S. to state mandated achievement assessments to the (nearly) infamous SAT (Scholastic Aptitude Test) and the ACT (American College Testing) college entrance exams.

This is an unfortunate situation, as most parents would much rather have their children learn for understanding rather than learn how to pass an examination. Further importance is placed on such examinations based on the simple fact that one must differentiate oneself if one wishes to qualify for the best possible opportunities.

We state that last comment in light of the fact that secondary education grades in the U.S. are highly inflated, see [98]. According to this article, 47% of graduating students in 2016 had an "A" average. An "A" average means that over the four-year course of study in a typical U.S. secondary school, 47% of students average an "A" across all courses taken, usually indicated by a minimum 90% scoring average in a given course. For further reference, the intention of the A-E grading scale is for a "C" to be the average grade, with "B" showing above average performance, and "A" reserved for excellent or outstanding performance.

The authors of this manuscript do not wish to opine as to the reasons for such inflation, only to note its existence and the additional reliance upon other standardized measures to aid university acceptance. If having excellent grades is not a differentiator, then we must rely on other measures, hence the added emphasis on the SAT and/or ACT in many cases. However, this also causes a problem.

A study cited by [1] states they looked at 123,000 students who enrolled in 33 colleges that do not require applicants to submit such test scores. According to the cited report, the differences in grade point average (GPA) for students who did not submit entrance exam scores and those who did was a paltry 0.05 (on a scale of 0.0 - 4.0), with an average of 2.83 compared to 2.88. Similarly, they show a 0.6% difference in graduation rates.

The fact that we have cited only two references aside, the issue is real. We have a system in which assessments (either self-report, supervisory, or academic) are required for us to advance in many facets of our lives. The reliability and validity of such assessments are suspect. We are told we

should not rely on their data and outputs by organizations such as the APA, and yet we apparently cannot get away from them.

5 APA Guidance on Evidence of Reliability and Validity

This section is intended to provide an outline of the American Psychological Association's (APA) guidance on gathering evidence of reliability and validity for assessments. We largely follow the APA three-volume series found in [75], [76], and [77]. We also attempt to follow the guidance found in [4]. We note that international organizations such as the British Psychological Society (BPS) and the European Federation of Psychologists' Association (EFPA) have nearly identical requirements. A future project is intended to provide a concrete mapping between the various requirements of each organization.

In Section 4, we lay out the issues we have with the current state of affairs in the assessment world related to our understanding of what is an acceptable level of evidence of either reliability or validity. We feel that the entire community would be greatly aided with more definitive guidance on the topic. Having said that, we do feel that TTI Success Insights is far ahead of the vast majority of our competitors in the market in this area. We base this on our reading of the available technical manuals and the content therein.

TTI Success Insights has a talented and experienced Research and Product Development team dedicated to the continued success of our products. We have established a continual improvement process and monitor our assessment performance on a nearly daily basis. This team collaborates with external research partners and routinely publishes articles in peer-reviewed publications. One of our team members is often invited to speak at national and international conferences on neurology. It is our intent to continue to set our standards as high as possible and continue to lead the industry with our research and development program.

5.1 Evidence of Reliability

The APA's first volume on testing and assessment, [75], contains a chapter on evidence of reliability, [74]. In this chapter, the authors outline four areas for which to gather evidence of reliability. We first note that there is often confusion surrounding this concept. We would like to briefly address some of these issues before continuing to the APA guidance.

In Section 4.1, we highlighted some areas that the authors in [44] and [96] point to as what they called misperceptions. These misperceptions are specifically related to the measure of internal consistency which we will discuss momentarily. The first misperception is that the so-called α (alpha) coefficient was developed by Cronbach. It is the understanding of these authors that Cronbach gave the coefficient the name α, but the development should actually be attributed to Kuder and Richardson, see [105], some 20 years earlier.

The next misperception is that α is equal to the level of reliability of an assessment. This is false. We need to think of α as a lower bound on the reliability of an assessment. The main idea here is that the true error associated with an assessment cannot be completely known. Since reliability is very closely related to the error associated with an assessment, the true reliability score for an assessment cannot be completely known.

Another misperception, and a topic of great annoyance, is that α should be greater than 0.70. The annoyance is that while we present a range of values that we follow, see Table 4.1, even the original author removed this table from his work. Unfortunately, we have yet to find a definitive statement on acceptable levels of α coefficient values.

A final misperception we present is that reliability and validity are synonymous. This is not even close to being true. The two are related in the sense that for an assessment to be valid, it must

first be reliable. However, that is as far as the relationship goes. To put it in to more technical language, we say that reliability is a necessary, but not sufficient requirement for an assessment to be valid. There are many other misperceptions in the assessment world that we leave for another time.

5.1.1 Alternate Forms Reliability

We begin our discussion with the forms of reliability we are likely to never use. The main reason for this is the level of difficulty and cost associated with measuring this form of reliability. Alternate Forms Reliability is just what it sounds like. This form of reliability consists of having multiple versions of an assessment and testing them on the same population and comparing the response patterns of said population.

It is the opinion of the authors of this manual that the benefit gained from any knowledge from this type process is far exceeded by the cost in dollars, time, and other resources. Other measures of internal consistency exist that provide far better information and are much more cost effective, especially with the continued advances in technology. The fact that one may compute the α coefficient on a data set of 100,000 or more respondents in seconds should be evidence enough of the failure of the alternate forms method to be a reasonable choice today.

5.1.2 Generalizability

Generalizability is not what it would appear based on the name. The name appears to imply that we are attempting to determine whether an assessment may be generalized to populations other than the population used to develop the assessment. The reality is that this form of measuring reliability is actually an attempt to better identify the sources of the errors.

In theory, we really like this approach. In practice, it is quite difficult to collect enough of the correct type of data to apply the methodology. We have attempted in some areas to use the approach. For example, in an earlier version of our Emotional Quotient technical manual we devoted a chapter to this topic.. The results were solid, yet we will likely not continue to use this methodology due to the aforementioned issues with collecting appropriate numbers of respondent data in appropriate area in order to make the analysis possible.

5.1.3 Temporal Consistency

The APA uses the term test-retest reliability. We prefer to call this concept temporal consistency. The main reason for this choice is that we prefer to not use the term *test*. The word *test* comes with a connotation that one is being graded on a right/wrong or correct/incorrect scale. Our assessments measure scales of various behavioral concepts such as observable behavior, motivation, emotional intelligence, and so on. There are no correct answers. We measure the location an individual falls on a spectrum or continuum. For example, in our behavior assessment, we measure how one approaches problems and challenges on a continuum with Direct on one end of the spectrum and Reflective on the other. An individual is either more direct or more reflective. There is no right/wrong in this setting.

For temporal consistency, we require a population of individuals who have taken an assessment multiple times within a reasonable time period. Multiple times usually means at least twice and reasonable time period is often thought to be at least 60 days but not more than 13 months. We want to ensure that there is enough time between sittings for the assessment that we are actually

measuring the constructs and not testing the memory of the respondents. We also want to ensure that we do not wait too long as an individual's style or motivation may be impacted by major life events that become more probable the longer one waits.

The measurement for temporal consistency is a simple correlation coefficient. We are not as interested in comparing one individual's score today against their score from a year ago. We intend to measure each individual's relative standing within the population overall. In other words, we wish to determine how close the current rank ordering is to the previous rank ordering of the same population. This means we consider the correlation between the two sets of scores.

5.1.4 Internal Consistency Reliability

When people ask the question "Is your assessment reliable?", this is most likely what they are asking about, whether they know this or not. Internal consistency reliability is what is measured by the so-called α coefficient. This is not the only measure of internal consistency and it is likely not the best, but it is by far the most popular. The popularity of the α coefficient is likely due to the relative ease with which it may be computed and the relative ease with which one may interpret the results. Most statistical software has built in functions which one may call with their relevant data and within fractions of a second have the results.

For most of our suite of assessments, we measure multiple variations of internal consistency. We do measure the α coefficient. We also consider inter-item correlation where appropriate. We also measure the so-called α-with-item-deleted statistic. In some cases we have measured the ω (omega) coefficient from McDonald, see [126]. The point of all this is that no single measure provides a complete picture of the overall assessment. No single measure should be used to make decisions about keeping or removing items from a scale. It is the compilation of a host of information that tells the complete story.

5.2 Evidence of Validity

Much like when someone asks whether an assessment is reliable, when the same question is posed about validity, the individual does not truly understand what they are asking. Most individuals posing these questions want yes or no answers to a question that does not have a yes or no answer. As an example, which of the five areas of validity the APA suggests is the person asking about? The answer is that most do not know that there are at least five categories. This individual also is likely not aware that whether an assessment is valid is constructed over years of gathering evidence for each of the areas suggested by the APA or other relevant body. One must construct a valid logical argument for validity, not merely say yes or no or compute a *validity coefficient.*.

The main question being asked about when discussing the validity of an assessment is whether the assessment is measuring what it intends to measure. Care must also be taken with the use of the term validity overall. There is a reason the APA refers to *evidence of* reliability and validity. It is the intent of our research program and our continual improvement process to monitor our assessments and continue to gather the relevant evidence of validity to enhance our response to the question of whether our assessments are valid.

5.2.1 Validity Evidence Based on Test Content

The APA presents its basic vision for validity evidence in [163]. The first concept presented in their current advice is validity evidence based on test content, which we will refer to as test content for purposes of brevity. Test content essentially asks the following questions:

- Has the assessment covered the entire domain of the intended content?
- Has the assessment introduced concepts that are not relevant to its intended purpose.

In order to address such questions, we first need to understand what the domain of definition ia. Phrasing it a different way, we need to understand the theory underlying the constructs being measured by the assessment. Finally, we need to understand the appropriateness of the assessment construction procedures. We discuss each of these concepts individually.

Rather than attempt to rewrite [163], we present some information directly from the source. First up is the concept of construct definition.

> Construct definition refers to the appropriateness of how the domain to be tested is described and specified. In evaluating the construct definition, one evaluates the test developer's general statements regarding what is being tested as well as the specifications that describe test content. [163], pp. 65-66.

The second concept is that of construct relevance. We again take the information directly from the source.

> ... validation focuses on whether each item on the assessment is relevant to the construct tested. [163], pp. 66.

In other words, are we presenting items that marry with the intended purpose of the assessment? A closely related concept is construct representation.

> ... construct representation ... investigates the degree to which the items and tasks on an assessment fully represent the intended construct and do not contain material irrelevant to the construct measured. [163], pp. 66.

Finally, we consider the appropriateness of test construction procedures.

> ... evaluation of the appropriateness of test construction procedures involves looking at the various item development, selection, scoring, and quality control procedures involved in constructing the assessment. [163], pp. 66.

We attempt to outline responses to these questions in various places in each of our technical manuals. We provide a history of the concept and the theoretical underpinnings of each. The specific details of our assessments are considered proprietary, so we do not directly address the construct relevance and representation issues in this manual. However, we believe that the presentation in this work provides the reader with the confidence that we have done our due diligence in these areas. Appendices Q and R provide our adaptation protocols and an outline of our development processes.

5.2.2 Validity Evidence Based on Relations to Other Variables

Validity Evidence Based on Relations to Other Variables is a long way of saying whether the assessment is actually giving us relevant information about something outside of the assessment. This area of validity evidence has been known under the names of concurrent validity, predictive validity, and criterion validity over the years. The current naming convention may be thought of synonymously with criterion validity or with the combination of concurrent and predictive validity.

> Although criterion-related validity evidence has been established as insufficient to fully support claims of validity for score interpretations and use, depending on the purpose of the test such evidence may be quite useful in building the overall validation argument. Validity evidence based on relationships between test scores and other variables extends beyond single test-criterion relationships and includes analysis of relationships of test scores with constructs that are expected to be positively related, negatively related, or unrelated. Confirmation of such theoretical relationships can reinforce the interpretations and uses that are intended to result from a score on a given instrument. [163], pp. 66-67.

This quote is instructive in that it is providing a view into one possible roadmap to an overall validity argument. What is also clear is that no single criterion related study is sufficient. We must continually seek out opportunities to conduct such studies and document the results. We also need to verify the theoretical relationships. In other words, we cannot just *spray and pray*. We must intentionally seek out studies that we *a priori* expect to have positive or negative relationships and confirm those hypotheses.

One way we are attempting to fill in this information is with the use of our demographics collection which includes the O*Net industry and job classification. We are establishing relationships between the self-reported job of respondents and their response patterns in our assessment data. We have established such relationships with varying degrees of success largely related to the amount of data we have been able to collect for a given job/assessment combination. To date, our most successful criterion relationship studies are for our Style Insights assessment data which, not so coincidentally, is our largest seller. The point of the fact that Style Insights is our largest seller is that we have the best opportunity to have large enough samples for largest number of O*Net classification with this assessment. Our approach and some results are presented in Section 12. We are also in progress developing a *jobs manual* for each of our major assessments. This manual will be posted on our website when it is available.

5.2.3 Validity Evidence Based on Internal Structure

Validity Evidence Based on Internal Structure, referred to from here on as internal structure, is likely the most accessible category of validity evidence. This is, essentially, a mathematical or statistical confirmation of the concepts referred to in Section 5.2.1. Specifically, we are interested in measuring the dimensionality of our assessment and confirming that our response patterns support the theoretical foundation laid out during the assessment construction.

Depending on the response format of a given assessment, measurement of internal structure may be accomplished in many ways. One may consider the concept of (corrected) item to total score correlation, which measures how well correlated each item is with the total score when that item is removed from the scoring. In this manner, we are testing to see if each item works well with the remaining items on the scale. Particularly low values may indicate additional and undesired dimensions may exist in the data. It may be the case that a certain construct contains multiple dimensions intentionally. In that case, one could measure the item to total score correlation for the sub-constructs individually.

Another possible approach is to use some form of item response theory. TTI Success Insights uses the graded response models of Samejima, see [155], when analyzing our Likert-style response format data. For our forced-rank response format data we employ the Thurstonian item response theory model from Brown and Maydeu-Olivares, see [35] among many others.

One may also employ a host of factor analytic approaches including exploratory factor analysis, confirmatory factor analysis, and structural equation modeling. Technically speaking, confirmatory factor analysis is a form of structural equation modeling, but we are arguing semantics at that point. We employ the exploratory factor analytic approach to analyze many of our assessments' response data. In the case of our Likert-style response data, we directly employ the factor analytic approach to measure the (sub-) construct dimensionality, analyze the factor pattern matrices, and estimate the variance explained by the scales.

For our forced-rank response format assessments, we employ the exploratory factor analysis to the individual scales and use the results for diagnostic purposes. We recognize that the tool may not be the most appropriate for the job, yet the research community has been slow to respond to the need for tools to analyze this type assessment. It is also recognized that the technique is not applicable to the full assessment for technical reasons. For those unfamiliar with the approach, factor analysis is sometimes referred to as the analysis of covariance. In the solution of the underlying mathematical problem one must invert the population covariance matrix. This is most easily understood as "dividing" for the mathematical layperson. The problem is that a forced-rank assessment generates a singular covariance matrix, by definition. A singular matrix cannot be inverted. This is the equivalent of trying to divide by zero.

There are also detractors from the use of forced-rank assessments in general, see [99]. We have read their concerns and are attempting to follow their advice regarding limiting our interpretation of the results of our analyses as well as attempting to find more appropriate methodologies such as the Thurstonian IRT modeling approach mentioned above. However, the Thurstonian IRT model has its detractors as well. The bottom line is we use techniques that may not be completely appropriate to the task, use them carefully for diagnostic purposes, and are completely transparent about it.

5.2.4 Validity Evidence Based on Response Processing

A description of this area of validity evidence is best read through the words of the APA themselves.

> Gathering validity evidence based on response processes is perhaps the most difficult validity evidence to gather because it involves demonstrating that the examinees are invoking the hypothesized constructs the test is designed to measure in responding to the test items. ... Gathering this type of evidence is difficult because one cannot directly observe the cognitive processes going on within people's heads as they respond to test items. [163], pp. 76.

Maybe we can. It is not the intention of this section to take away from other sections of this manual. Section 13 is written by our resident neurological expert. He gathers electroencephalograph (EEG) images of the brain as examinees respond to our assessment items. In this way, we have the ability to confirm that the assessment taker is invoking the desired constructs when responding to our questionnaires.

There are other methodologies that we could explore. For example, we naturally measure item difficulty, either directly or through the use of one of the IRT methodologies.

> Similar to evidence-centered design, some educational test specialists have investigated modeling item difficulty on the basis of specific attributes of test items related to cognitive processes. ... When item difficulty can be explained using the cognitive attributes of the items, validity evidence based on response processes is provided. [163], pp. 78.

It is our intent to explore extending our approach to capture this idea of relating item difficulty to the cognitive attributes of our assessment items.

5.2.5 Validity Evidence Based on Consequences of Testing

Validity evidence based on consequences of testing consists of measuring the intended and unintended consequences that may arise from the use and interpretation of assessment scores and data. Further, in our opinion, this includes constructing and implementing mitigation approaches to help lessen any negative consequences that may exist. For example, TTI Success Insights has a large network of distributors that we cannot possibly monitor on any consistent basis. We identify this as a potential negative consequence. Our mitigation strategy includes providing multiple sources of training and certification, both in-person and online, to help ensure that our tools are used as they are intended. We also ensure our internal customer service representatives are trained to spot potential misuses of our tools such as attempting to use them for clinical purposes or attempting to use them with young children.

One of the biggest challenges we face as an organization is that many of the policies and procedures we have were designed with the best of intentions but without knowledge of concepts such as gathering evidence of validity based on consequences of testing. In other words, we have many mitigation strategies in place, but these processes and why they exist have not been completely documented over time. We are working to identify as many as possible and get them properly documented. See also Section 14 for more on this topic.

5.3 Deductive Versus Inductive Reasoning

Human beings are hard wired to observe. It is a basic survival functionality within us. We also know that we do not always see what really happened. The Constitutional Rights Foundation https://www.crf-usa.org/ states on its website that approximately half of all wrongful convictions are due to mistaken eyewitness testimony. So, what's the point? The point we want to discuss is the difference between deductive and inductive reasoning, why one of them is logically valid and the other is not, and how that applies to our situation in the assessment community.

Suppose an alien comes to visit planet Earth and ends up visiting a small rural town somewhere in the United States. The town only contains people with blue eyes. This being the only humans the alien has ever seen, the alien is tempted to conclude that all humans have blue eyes. Of course, we all know that there are many shades of eye color, so any inference drawn from the population of that small town is false. This is an example of inductive reasoning.

Inductive reasoning is essentially a bottom-up thought process. We observe the world around us, we use those observations to draw conclusions about the world, and we are all too often wrong. Why? The biggest reason we are often wrong is that we fail to recognize that we are reasoning inductively and we can at best estimate the probability of something being true. We confuse the probability that a given thing may be true with an absolute conclusion that it is true. If even one person on the entire planet of more than 7 billion people has an eye color different from blue, our friend the alien's conclusion in the previous example is false.

This is not to say that inductive reasoning is pointless. In fact, many great discoveries in our scientific community have come from inductive reasoning, at least to start. Copernicus studied the motion of the planets and came up with the theory that our solar system was heliocentric. However, he could not prove his theory with inductive reasoning. Similarly, Kepler studied the motion of the planets and formulated the idea that the orbits are elliptical, not circular. For those

who desire precision, yes, a circle is an ellipse with eccentricity equal to one. He could also not prove his assertions with observations. It takes something more because, as previously noted, even one case of an observed phenomenon that fails to adhere to any conclusions drawn about the observed phenomenon leads to a false argument.

Deductive reasoning, on the other hand, is a top-down thought process. We start with a premise (a theory, an assumption, etc.) and we follow a series of logical steps to a conclusion. The classic example of deductive reasoning presented in most introductory courses on philosophy or logic is as follows. All humans are mortal. Socrates was a human. Therefore Socrates was mortal. The premises are "humans are mortal" and "Socrates was human", the conclusion is "Socrates was mortal".

There are many differences between inductive and deductive reasoning. The largest difference may be characterized by saying that deductive reasoning is logically valid while inductive reasoning is not. In the deductive case, if the premises are true and the conclusion is true, they are true for all such cases that arise. In the previous example, the statements "humans are mortal' and "Socrates was human" are always true. The conclusion "Socrates was mortal" is also always true. We do not need to compute the probability that "Socrates was mortal". However, in the inductive reasoning example, we may be tempted to incorrectly assume that all humans have blue eyes. According to the website https://www.aucklandeye.co.nz/, blue eyes are genetically recessive and only about 8% of the world's population fall into this category.

As noted in the introductory paragraph of this section, human beings are hard wired to observe the world around us. We are also inclined to attempt to generalize and abstract. How do we come to terms with all these competing interests? In reality, we do both. We use our powers of observation to see what is happening in the world. We then process that information and attempt to generate theories about how it all works. Once we have our theories, we attempt to prove that they must always hold. So, we have a tendency to reason inductively through our observations and then attempt to reason deductively to prove theories about what we observe. It is an infinitely long chain of iteration between the two.

5.4 Summary

In some sense, the previous arguments related to the various forms of reasoning are quite applicable to the world of assessment. Many assessments fall into the category of having been developed through many years, decades, and more, of observing human behavior. The TTI Success Insights Style Insights is an assessment that is often described as measuring an individual's observable behavior. In that sense, we may want to refer to it as being based on inductive reasoning. On the other hand, the TTI Success Insights Acumen Capacity Index, which is based on the Hartman Value Profile, is developed based on the theory of value. In that sense, we may wish to refer to this assessment as being based on deductive reasoning.

The difference is a subtle, yet important one. We interpret this to mean that in the case of inductive reasoning based assessments, we have a larger burden of proof relative to the categories of evidence of reliability and validity outlined in Sections 5.1 and 5.2.

> If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the **user** to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary. [4], pp. 24.

Since many of our assessments are founded in the area of inductive reasoning, we are truly dealing

with a probability of a given thing being true. Hence, the quote above is directly stating that any time we encounter something we haven't used the assessment for, we must validate the use of the assessment in that situation. Alternately, if we have an assessment founded in deductive reasoning, the results are true always, ***assuming*** a valid, logical proof has been provided that the conclusion always follows from the premise.

We finish with a simple example to show how we may begin with inductive reasoning, develop a hypothesis, then finish with deductive reasoning to prove (or disprove) the hypothesis.

Example 5.1. *The exercise is to find a closed form for the sum of the first n integers. We begin by writing out the first few sums for $n = 1, 2, 3, \ldots$ The first of these sums is given by*

$$\sum_{i=1}^{n=1} i = 1 = \frac{1(1+1)}{2}. \tag{5.1}$$

Next, for $n = 2$ we have

$$\sum_{i=1}^{n=2} i = 1 + 2 = 3 = \frac{2(2+1)}{2}. \tag{5.2}$$

Continuing

$$\sum_{i=1}^{n=3} i = 1 + 2 + 3 = 6 = \frac{3(3+1)}{2}. \tag{5.3}$$

A final example is

$$\sum_{i=1}^{n=4} i = 1 + 2 + 3 + 4 = 10 = \frac{4(4+1)}{2}. \tag{5.4}$$

So, inductively we have observed that the sum of the first n integers appears to be related to the product of n and $n+1$ (apparently divided by 2). So, we create the hypothesis, stated as a ***theorem****, as follows.*

Theorem 5.2. *Fix a positive integer n. The sum of the first n integers is given by the following formula:*

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}. \tag{5.5}$$

Now, the deductive step is to prove (5.5) is true $\forall n \in \mathbb{Z}^+$. Note that $\forall n \in \mathbb{Z}^+$ should be read **for all positive integers** n. In some works the set $\mathbb{Z}^+$ may be called the natural numbers and would be denoted $\mathbb{N}$.

Proof. The easiest way to prove this is to use the technique known as mathematical induction, which should not be confused with the term inductive and its use earlier in this section. The steps are to first prove that (5.5) is true for $n = 1$. We then assume the statement (5.5) is true for an arbitrary k and use this to establish the statement is true for $k + 1$.

The case for $n = 1$ is trivial since the sum of the first positive integer is 1, so

$$\frac{1(1+1)}{2} = \frac{2}{2} = 1. \tag{5.6}$$

We now assume (5.5) is true for an arbitrary k, so assume that

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2} \tag{5.7}$$

is true. Then, we prove the statement is true for the next positive integer $k+1$.

$$\sum_{i=1}^{k+1} i = \sum_{i=1}^{k} i + (k+1) = \frac{k(k+1)}{2} + (k+1) = \frac{k^2+k+2k+2}{2} = \frac{k^2+3k+2}{2} = \frac{(k+1)(k+2)}{2} \tag{5.8}$$

as desired. QED. □

6 Evidence of Reliability Based on Internal Consistency

This section presents the results of internal consistency reliability analyses of the TTI Success Insights Emotional Quotient (EQ) assessment. The analysis focuses on the more commonly reported coefficient α analysis as well as extending these analyses to include coefficient α with confidence intervals, α-if-item-deleted, and inter-item correlation analyses. This section begins with an overview of reliability and internal consistency as a measure of reliability, following the guidelines from the viewpoint of the APA, set forth in [74].

The APA makes some significant comments regarding reliability and its relationship to psychometric assessments. First, the APA loosely defines reliability as follows, see [75], chapter 2, page 21:

> To the extent that the scale is subject to such random errors, it is said that the scale is not reliable.

In the quote above, random error refers to errors that are considered to be random because people do not know or choose not to know their source. It is not the case that we are in the setting of quantum physics where spooky action at a distance guides the theory, at least according to a paraphrased description once given by Einstein regarding the then new ideas of quantum theory. Instead, it is acknowledged that random errors exist and a choice is made to measure the error rather than seek out the source of the error. After all, if the error associated with a given thing is small enough, we may disregard said error.

The APA also makes the following statement, again in [74], chapter 2, page 22:

> Validity is more important than reliability to be sure, but reliability is generally considered to be one of the most important criteria against which measures, and more importantly, scores that result from them are evaluated.

Finally, we present two quotes that are intended to show the relationship between reliability and validity from the viewpoint of the APA. The first is regarding what it means for a test to be completely unreliable, [74], chapter 2, page 23:

> ... a completely unreliable test can measure nothing because the numerical scores that are assigned to the examinees are essentially random numbers and will not bear a consistent relationship to any other attribute among the individuals being assessed.

Consider also the following from [74], chapter 4, page 61:

> ...[Validity] refers to the degree to which an assessment fulfills its intended purpose and test results are appropriately interpreted.

Combining the two thoughts, it appears that if an assessment cannot be considered reliable, it cannot be valid. If the numerical scores based on an assessment do not bear a consistent relationship to any other attribute among those being assessed, the assessment cannot be fulfilling its intended purpose. Put another way, an assessment cannot be valid if it is not first reliable. The question may be posed in reverse. Can an assessment be reliable if it is not first valid? The answer to this question is yes.

The authors of this manual have used the example of a home oral thermometer in the past. Consider such a thermometer. If the thermometer consistently measures the same temperature under the same or similar circumstances, then it is reliable. If the thermometer readings are also correct, it is then considered valid. A slightly more technical way to state this is reliability is a necessary, but not sufficient, condition for validity.

6.1 A Short Review of the American Psychological Association's View of Reliability

Generically speaking, reliability is possibly best described as consistency. Consistency may be measured across time, across similar instruments, across content sampling, or across multiple *facets* other than time and content sampling. These are the four main areas of reliability the APA chooses to concentrate on in [74]. In more common language, the first three of the concepts just mentioned refer to test-retest reliability, parallel or alternative forms reliability, and internal consistency. The fourth concept refers to an area called Generalizability Theory or G-theory for short. The reader interested in more information on G-theory may consult, among many others, [60] for the original work on the topic and [178] for a brief introduction.

Test-retest reliability, also referred to as temporal consistency, measures reliability with respect to time. In other words, a test-retest approach to reliability measures the responses of individuals to an assessment taken at two different times and compares the results. See Section 7.1 for a more detailed explanation. Alternate forms reliability also requires multiple assessments to be administered. In alternate forms, two forms of an assessment are generated that purport to measure the same content domains. The assessments are administered and results compared, see [74], pages 30-32, for more on alternative forms reliability.

The main focus of the remainder of this section is internal consistency reliability of the TTI Success Insights EQ assessment.

6.2 A Brief Discussion of Coefficient α

The α coefficient has, in recent years and perhaps decades, been the most widely reported measure of internal consistency. There are many likely reasons for this, first and foremost being the relative ease with which the α coefficient may be calculated. From an interpretive standpoint, α is a scaled version of item correlation.

The myth that coefficient α was first introduced by Lee J. Cronbach in his work [58] in 1951 persists today. The reality is that the α coefficient first appeared in [105] in 1937. Contemporary understanding of internal consistency has evolved and more robust measures of internal consistency do exist, yet α persists in both the academic and business communities. Section 6.4 of this manual discusses the relationship of internal consistency measures such as coefficient α and item correlations. For more in-depth discussions, see any of [20], [139], [146], [167], [169], and [179] among many others.

Many discussions have occurred over the years as to what is acceptable internal consistency and how to best report it. Cronbach himself suggests that the α coefficient is not the best measure of internal consistency and suggests better measures exist, see [59]. The authors mentioned in [20], [139], [146], [167], [169], and [179] offer alternatives, among which are reporting α coefficient with confidence intervals as presented in Section 6.5. Additionally, coefficient ω, with confidence intervals is also discussed in the aforementioned works.

6.3 α-if-item-deleted as a Measure of Internal Consistency

The analysis known as α-if-item-deleted is a rather straightforward concept. Sometimes this technique is called α-without. The basic idea is to remove a particular item and compute the α coefficient with all other items retained. In theory, if the item is consistently contributing to the reliability of the instrument, removing it will have an adverse impact on the reliability of the instrument. This

will be seen as a lowering of the α coefficient. Otherwise, the α coefficient will increase, indicating a potential problem.

In most cases, any increase in the α coefficient indicates that an item is not functioning as intended. However, any increase is not necessarily cause for removal of the item from the instrument or removal of the item from consideration if one is building a new questionnaire. As is noted throughout this manual, it is the sum total of information that should be considered before making final decisions on an item, whether to keep or to remove. A small increase in the α coefficient when an item is removed could be caused by inconsistency elsewhere in the instrument. One should further consult the inter-item correlation statistic presented in the next section, along with other measures such as corrected item-to-total score correlation, see Section 11.1, and exploratory factor analysis, see Sections 10 and 11.3.

6.4 Inter-item Correlation as a Measure of Internal Consistency

Inter-item correlation is a measure of the correlation between the item responses. It is computed using the standard correlation coefficient based on the item level responses to the items of the assessment. Average inter-item correlation for an item is merely the average of all non-redundant correlations of this item with all other items being measured by the assumed scale. What is meant by non-redundant is that we only consider the covariances and remove the variance from the computation of the mean. In other words, if a scale has n items, the average inter-item correlation for item j is the average of the $n-1$ correlations between item j and items $k = 1, ..., n$, $k \neq j$.

The literature is a bit vague on appropriate thresholds. The two most popular ranges for the average of all inter-item correlations for a construct appear to be $0.15 - 0.50$ and $0.20 - 0.40$. The only range TTI Success Insights has seen concerning the individual items is $0.15 - 0.85$. However, the interpretation of a correlation between two variables with a correlation of 0.85 is that they are essentially the same. In the end, analysis of items and whether they are adequately measuring the construct is not based solely on one statistical test. Rather, an overall assessment of multiple statistical measures along with the eye-test need to be employed before making a final decision.

The analysis of this section focuses on two main areas, essentially the mean and standard deviation of the inter-item correlations. We are interested in the mean falling in the ranges mentioned in the previous paragraph, while at the same time having as little variation as possible across the individual inter-item correlations, see [45] for an in depth discussion of the reasoning. In short, a good mean addresses the internal consistency aspect of reliability. Small standard deviation leads more into the aspects of validity which is discussed in Sections 10 and 11.1.

The authors in [45] present a quote of B. F. Green that articulates well the idea we are attempting to describe. See [90] for the original reference.

> Put another way, to ensure uni-dimensionality, almost all of the inter-item correlations should be moderate in magnitude and should cluster narrowly around the mean value. B. F. Green articulated this principle most eloquently, stating that the item intercorrelation matrix should appear as a "calm but insistent sea of small, highly similar correlations." [45], pp. 316.

The myriad measures of internal consistency are quite closely related. For example, it may be shown that the α coefficient presented in Section 6.2, first introduced in [105], is a scaled version of the average inter-item correlation. If we denote the average of all non-redundant correlation coefficients

by $\bar{r}$ and assume there are K components of a scale which we are measuring, one estimate of α, called the standardized estimate, is given by

$$\alpha_{se} = \frac{K\bar{r}}{(1 + (K-1)\bar{r})}. \tag{6.1}$$

The representation of α_{se} in (6.1) shows that the standardized α coefficient is a scalar multiple of the average inter-item correlation coefficient. For the reader interested in more on this topic, consult any of [20], [139], [146], [167], [169], and [179], among many other possibilities.

Of particular note, it is argued in [169] that average inter-item correlation does not reflect the variability among the inter-item correlations. In agreement with this line of reasoning, it is the position of TTI Success Insights that a single measure of reliability and/or internal consistency is not enough information to make fully sound decisions. As discussed further throughout this work, multiple measures are needed. This is consistent with the findings of [169], although the current analyses divert from the details of the information suggested in that article.

6.5 Internal Consistency Results

For interpreting reliability coefficients in general, Table 6.1 is presented. An exact source for the table is difficult to track down. It appears that a version of this was originally published in the first edition of [63]. However, on page 184 of [12], the author states:

> How high should values of α (or related indices such as KR-20) be? I hesitate to give an answer to this question because estimates of α can be affected by many things, such as the purpose of the test, the heterogeneity of the sample from which it was obtained, the conditions of testing, and the number of items.

Table 6.1: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

[12] goes on to present several examples of the differing opinions on what constitutes acceptable levels of evidence of reliability based on the α coefficient or similar measure. For example, [130] states that 0.70 is acceptable during development, 0.80 for "basic research scales". and 0.90 for clinical settings are appropriate minimums. [144] recommends a minimum of 0.80 without the same restrictions as [130]. As a final example, [101] suggests values of 0.95 for clinical settings. [12] concludes her remarks on recommended levels of reliability with:

> The higher values for clinical decisions reflect the serious consequences of such decisions and the need for correspondingly higher standards. The same argument for higher reliabilities could be made in the context of educational tests that are used for student

> placement, graduation, or remediation decisions or for licensure or certification tests. Indeed, any test that will be used to make consequential decisions about students, clients, patients, or others, should contain as little error as possible.

It is beyond the scope of this document to argue the merits of Table 6.1 or the merits of any particular bounds that have been discussed over the years. We present Table 6.1 along with the comment that some authors suggest any reliability coefficient with value exceeding 0.70 to be acceptable while others suggest much more conservative estimates. The author in [63] states unequivocally that these bounds are his preference and that "I cannot defend them on strictly rational grounds."

We first present a breakdown of information in the TTI Success Insights general population[3]. We note that the smaller sample size noted in Table 6.2 is due to making adjustments to the underlying questionnaire. These adjustments have increased the overall evidence of internal consistency reliability and addressed evidence of validity in several areas. To see the evidence of improved internal consistency, compare Tables 6.3 and A.7 (inter-item correlation), Tables A.10 and A.12 (α coefficient), and Tables A.15 and A.15 (α-if-item-deleted) below.

The evidence of validity improvements may be seen in Section 8 (test content validity) and by comparing the results in Sections 11.4 and B.1 (internal structure validity). Our historical development is outlined in Section 8.5.

Table 6.2: U.S. General Population Statistics Standard Scoring
N = 4,395

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	72.36	13.98	71.27	14.12	73.05	13.74
Self Regulation	68.75	12.98	70.80	12.72	67.11	12.98
Motivation	74.37	13.88	71.18	13.75	74.26	13.90.
Social Awareness	71.07	12.47	68.29	12.82	74.28	11.18
Social Regulation	70.13	12.40	70.13	13.47	70.19	13.12

6.5.1 U.S. Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the English-U.S. version. The results are presented in Table 6.3. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability. The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the U.S. version based on inter-item correlation.

[3]The information presented in Table 6.2 is based on our first look at the data based on an updated questionnaire, hence the relatively small population

Table 6.3: Average U.S. Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.46	0.35	0.37	0.28	0.39
Item 2	0.43	0.36	0.37	0.27	0.37
Item 3	0.41	0.37	0.37	0.28	0.36
Item 4	0.44	0.35	0.36	0.27	0.36
Item 5	0.41	0.35	0.37	0.28	0.38
Item 6	0.44	0.37	0.37	0.29	0.37
Item 7	0.41	0.36	0.38	0.31	0.36
Item 8	0.44	0.35	0.37	0.28	0.37
Item 9	0.41	0.35	0.37	0.28	0.36
Item 10	0.43	0.35	0.36	0.26	0.36
Item 11		0.36	0.37	0.26	0.37
Item 12		0.35	0.37	0.29	
Mean	0.43	0.36	0.37	0.28	0.37
StDev	0.02	0.01	0.00	0.01	0.01

6.5.2 U.S. Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the English-US version. The results are presented in Tables A.10 and A.15. Table 6.1 may be consulted for the relevant scoring categories when reviewing the data in Table A.10.

We are presenting two variations of internal consistency. The first is presented in Table A.10 and is the measure of internal consistency based on the α-coefficient and presented with confidence intervals. The second is presented in Table A.15 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values for the α-coefficient are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table 6.4: Average U.S. α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.88	0.86	0.87	0.81	0.86
Lower CI	0.88	0.86	0.87	0.81	0.86
Upper CI	0.88	0.87	0.87	0.82	0.87

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

There are no items for which the value of the α-coefficient increases in this analysis. There are a total of 7 items for which there is no decrease which is worth at least noting.



Table 6.5: α-if-item-deleted for U.S.

Est./Scale	SA	SR	MO	SoA	SoR
α	0.88	0.86	0.87	0.81	0.86
Item 1	0.88	0.85	0.86	0.80	0.86
Item 2	0.87	0.85	0.86	0.80	0.86
Item 3	0.86	0.86	0.86	0.80	0.85
Item 4	0.88	0.85	0.85	0.79	0.85
Item 5	0.86	0.85	0.86	0.80	0.86
Item 6	0.88	0.86	0.86	0.81	0.85
Item 7	0.86	0.85	0.87	0.82	0.85
Item 8	0.87	0.85	0.86	0.80	0.86
Item 9	0.86	0.85	0.86	0.81	0.85
Item 10	0.87	0.85	0.86	0.79	0.85
Item 11		0.85	0.86	0.79	0.85
Item 12		0.85	0.86	0.82	

6.6 Summary

The results presented in Section 6.5 above show five constructs with very solid evidence of internal consistency reliability. This is confirmed whether we are looking at the traditional α coefficient presented in either Table A.10, where we present the coefficient values with the corresponding confidence intervals, or in Table A.15, where we present a variation of the α coefficient that is computed during the α-if-item-deleted analysis. The variants are quite consistent in their values. Further, we see in Table 6.3 very solid values of the inter-item correlation statistic. These values confirm the high levels of internal consistency.

As is noted elsewhere in this document, as well as in [75], establishing the reliability and validity of an instrument is not a one and done analysis. Reliability and validity of psychometric instruments are established by showing the relevant measures hold up statistically over time. For this reason, we present the historical information in Appendix A. Table A.7 shows the analysis from the most recent inter-item correlation analysis prior to the aforementioned update to the questionnaire. Tables A.12 and A.17 show the relevant historical α coefficients and α-if-item-deleted analysis, respectively. All relevant historical information is presented in Tables A.7 through A.14. Further, additional historical information may be provided upon request.

7 Evidence of Reliability Based on Temporal Consistency

This section discusses some of the underlying theory of temporal consistency as well as presenting results from selected test-retest studies conducted by TTI Success Insights. The organization of this section is as follows. First, we present the APA's view of test-retest reliability, followed by rest-retest study results.

7.1 A Brief Review of the APA's View of Temporal Consistency

Temporal consistency is one of the four main areas of reliability that are endorsed by the APA, see [75] chapter 2 on Reliability. According to the author of this chapter, [74], reliability in the sense of temporal consistency is defined as stability. In other words, an assessment may be considered reliable if assessment scores are stable over time. It is necessary to be a bit more explicit about what is meant by assessment scores being stable over time.

According to [74], page 28,

> One should note that stability of scores is defined as persistent relative standing within the group rather than consistent absolute value of score.

We interpret this statement to mean that it is to be expected that some differences may exist in scores on application of the same assessment over time. What is truly important is that the relative standing of the individuals is generally expected to remain consistent across time. At the end of the day, this is a lot of work to say that the correct tool to use is the standard Pearson correlation coefficient to measure the differences between individuals' scores from time 1 to time 2.

A major concern of the test-retest methodology is whether the quantity being measured is the reliability of the assessment or a measure of the memorization skills of the assessment respondents. The author of [74] states that there are (at least) four factors that memory may exert influence on and impact measurement of evidence of the reliability of an assessment. These are:

1. The length of time between administration of the assessments
2. The length of the test
3. The nature of the test materials
4. The nature of the characteristic(s) being measured

We address each of these topics one at a time. For the length of time between administrations of an assessment, it is clear that if the time between assessments is too short, there is a higher likelihood of memorization of items. This would then potentially inflate estimates of reliability. The length of time should be long enough that memorization becomes unlikely while not so long as to have major life changes occur. This assessment is measuring emotional quotient and we expect that over our lifetimes, our emotional intelligence changes (improves). Waiting for an exceedingly long time between administrations would likely cause an underestimate of the reliability of the assessment.

The only guidance the APA provides on the subject is as follows, see [74], page 29:

> ... the time period between the test and the retest should correspond to the length of time between typical testings and the subsequent behavior that the test attempts to predict.

It is not clear that the TTI Success Insights EQ assessment is predicting behavior in the sense mentioned in the above quote. What should be clear is that the main concern here is the length of time between assessments be such that memorization is unlikely. Further, one may be able to implement procedures intended to identify potential abuses of the assessment.

The current length of the TTI Success Insights EQ assessment is 57 items. These items are split along five scales and the items are randomly placed throughout the assessment. It is the opinion of TTI Success Insights that the combination of an adequate elapsed time between administrations of the assessment and its current length and makeup are sufficient to address the majority of rote memorization techniques.

The 57 items on the TTI Success Insights EQ assessment are currently scored on a 6-point Lickert scale. We again note that the length of time between administrations of the assessment and the length of the assessment discourage memorization. We do note that while it is possible to memorize a sequence of 57 numbers from the set of the first five positive integers, we feel it is unlikely that a large enough proportion of any test-retest population has the motivation to do so.

Finally, the nature of the characteristics being measured is addressed as follows. This assessment is measuring emotional intelligence. It is possible that memorization becomes a problem, but TTI Success Insights feels this is unlikely. Given that this assessment is measuring emotional intelligence, it is more likely that the state of emotions of the individual at the time of administration will directly impact the reliability of the assessment and that memorization issues are not a major problem for TTI Success Insights EQ.

A second area of concern related to temporal consistency is called the differential practice effect. Differential practice effect is the concept that performance, generally speaking, improves with practice.

According to the APA, see [74], page 29:

> Unfortunately, in many testing situations, different examinees have varying degrees of test-taking experience that may result in unequal performance increments between testings. In that case, the test-retest reliability may not be completely appropriate.

As mentioned in the note above, if we have an adequate response to the time issue, this is likely more than enough to address the differential practice effect. The fact that TTI Success Insights assessments (in particular EQ) are not measuring performance implies that as long as the respondents are not overly familiar with the assessment itself, we should be adequately addressing any issues related to differential practice effects.

7.2 Test-Retest Studies

TTI Success Insights regularly conducts studies into all facets of evidence of reliability and validity of our assessments as part of our ongoing continuous improvement process. Given that the TTI Success Insights EQ assessment is relatively new and is an as yet growing portion of the overall suite of TTI Success Insights assessments, the amount of data available for test-retest comparisons is rather small.

The current test-retest iteration is based on a total population of 1,222 with 260 respondents self-reporting Male, 234 self-reporting Female, and the remainder responding N/A. We consider records that indicate an individual who has taken the TTI Success Insights EQ assessment more than once

during a period of not less than 2 months and not more than 24 months. The relevant data are presented in Table 7.1

Table 7.1: 2024 Test-Retest Data Breakdown
TTI Success Insights EQ

	4/2/2018 to 4/10/2023
Elapsed Time	2-24 Months
Mean Elapsed Time	319 Days
Median Elapsed Time	276 Days
Population†	1222
Males†	260 (43%)
Females†	234 (57%)

Note 7.1. † *Approximately 69% of respondents answer N/A to the demographics question related to gender.*

Table 7.2: 2024 Test-Retest Correlations
TTI Success Insights EQ

Scale	SA	SR	MO	SoA	SoR
SA	0.62				
SR		0.71			
MO			0.71		
SoA				0.74	
SoR					0.75

Table 7.2 contains the correlation results between the first and second administrations of the TTI Success Insights EQ assessment in the data from April 2, 2018 to April 10, 2013. The results are solidly in the 0.70 to 0.80 range with all correlations computed with statistical significance at less than a 0.001 level. The exception is the SA scale coming in with a test-retest correlation of 0.62. It is not clear why this particular scale scores significantly lower than the remainder of the scales. However, it is noted that a correlation above 0.50 is solid in this setting and, while the TTI Success Insights continuous improvement process is in place to correct issues such as this, there is no major concern at this time.

7.3 Test-Retest Summary

The preceding section has presented an overview of the test-retest process as a measure of reliability following the guidelines of the APA as set forth in [74]. The results of the most recent version of the test-retest analysis, presented in Table 7.5, show solid evidence of test-retest reliability for the time period from 4/2/2018 to 4/10/2023. While the sample is somewhat small at only 1,222 total respondents, the results show strong correlation in 4 of the 5 scales and an acceptable level of correlation in the final scale.

7.4 Historical Test-Retest Results

This section contains the historical temporal consistency results.

Table 7.3: 2018 Test-Retest Data Breakdown
TTI Success Insights EQ

	1/1/2011 to 12/14/2014	12/16/2014 to 5/7/2018
Elapsed Time	2-24 Months	2-24 Months
Mean Elapsed Time	319 Days	352 Days
Median Elapsed Time	276 Days	324 Days
Population	203	705
Males	88 (43%)	403 (57%)
Females	115 (57%)	302 (43%)

Table 7.4: Test-Retest Correlations 1/1/2011 to 12/14/2014
TTI Success Insights EQ

Scale	SA	SR	MO	SoA	SoR
SA	0.56				
SR		0.75			
MO			0.75		
SoA				0.74	
SoR					0.73

Table 7.5: Test-Retest Correlations 12/16/2014 to 5/7/2018
TTI Success Insights EQ

Scale	SA	SR	MO	SoA	SoR
SA	0.65				
SR		0.72			
MO			0.70		
SoA				0.78	
SoR					0.75

8 Evidence of Validity Based on Test Content

We discussed the main areas related to validity evidence based on test content in Section 5.2.1 above. In the next several sections, we present a short refresher on each of the areas that the APA suggests are the relevant areas of evidence that is necessary to establish test content validity. Each section also discusses what TTI Success Insights is doing to satisfy the APA requirements. We begin with the area known as construct definition.

8.1 Construct Definition

Since we have already presented these concepts above in Section 5.2.1, we present a condensed version of the definition. For a comprehensive treatment, we refer the reader to the aforementioned Section 5.2.1 or to the original source [163] where the APA presents a lengthy and worthy discussion of this concept.

> Construct definition refers to the appropriateness of how the domain to be tested is described and specified. In evaluating the construct definition, one evaluates the test developer's general statements regarding what is being tested as well as the specifications that describe test content. [163], pp. 65-66.

At the end of the day, the point of construct definition is to ensure that what is being measured is clearly and adequately defined. A very pointed quote related to adequate definitions is provided by Michael Spivak in his short, but very powerful text on Calculus on Manifolds.

> There are good reasons why the theorems should all be easy and the definitions hard. As the evolution of Stoke's Theorem revealed, a single simple principle can masquerade as several difficult results; the proofs of many theorems involve merely stripping away the disguise. The definitions, on the other hand, serve a twofold purpose: they are rigorous replacements for vague notions, and machinery for elegant proofs. ... Concentrating the depth of a subject in the definitions is undeniably economical, but it is bound to produce some difficulties for the student. pp. ix, [166].

Clearly, most of us are not math students, and most of us will never have the need to understand Stoke's Theorem, the history, or its proof. However, the power of Spivak's viewpoint is that in understanding any subject, we must begin with the definitions. Moreover, if we do our hard work in properly defining the relevant areas of the subject, we find our conclusions much easier to come to. Perhaps a simple example is in order. Suppose we are asked to compute the area of a triangle. It impossible to do so without first understanding what a triangle is. For simplicity, let us assume our triangle is a right triangle. But what does that mean?

Definition 8.1. *A right triangle is a triangle with one of the three angles having a measure of* 90° *or* $\frac{\pi}{2}$ *radians, depending on your preference for degrees versus radians.*

The main reason for the assumption of a right triangle is to shorten the discussion. Every right triangle is exactly one half of a rectangle and we should all recall how to compute the area of a rectangle. For completeness:

Definition 8.2. *A rectangle is a parallelogram (4 sided geometric object whose opposing sides are parallel) such that each of the interior angles has a measure of* 90°*. The area of such an object is simply the product of any two adjacent sides.*



Well, that got sticky very quick. We now had to introduce the definition of *parallelogram* in order to define a rectangle. We also have *interior angle*, *adjacent sides*, and *product*. We could go on and on. Anyone who took a high school course in geometry will have seen all of these definitions, either in that course or in an earlier course, but do we all remember these definitions? Unlikely. To simplify, any product of two numbers is the area of a rectangle with side lengths equal to the two numbers we are multiplying together.

This may seem silly to some, but at this point we are ready to define the area of our right triangle (really any triangle). As noted above, every right triangle is half a rectangle and we know that a rectangle has area equal to its length multiplied by its width (the product of any two adjacent sides). We usually denote length by l and width by w. Hence, if we denote area by A we have

$$A = l \times w. \tag{8.1}$$

This directly leads to the area of a triangle as

$$A = \frac{1}{2} \times l \times w. \tag{8.2}$$

Since mathematicians and math teachers around the world like to confuse students, we usually write this as

$$A = \frac{1}{2}bh \tag{8.3}$$

where b denotes the base of the triangle and h the height of the triangle. We have also modified our notation of multiplication, removing the use of $\times$ and going with the traditional notation that bh really means $b \times h$. We have not introduced a single number into the discussion. We have used definitions to arrive at the solution of the problem of finding the area of a right triangle by putting the hard work into the definitions. We now reap the benefits of never needing to do this again as long as we can recall the simple formula in (8.3).

Getting back to the topic at hand, most of the TTI Success Insights EQ model is discussed in two main sections of this work. First, Section 2 gives a high level overview of EQ in general, including some of its history along with presenting the models of Bar-on, Goleman, and Mayer and Salovey. Of particular interest are Sections 2.3 and 2.4 which provide an in-depth discussion of Goleman's model and a presentation of a compare and contrast of the various models being discussed in Section 2.

Second, Section 3 provides an overview of the TTI Success Insights EQ model. Generally speaking, our approach is a hybrid, blended implementation of the Goleman model. There are a multitude of reasons why this is currently the case. We outline below in Section 8.4 a partial explanation for some of the reasons. It is important to notate some of the historical markers and, without naming names, the individuals involved.

In the late 2000s, TTI Success Insights came into an opportunity to purchase the rights to an emotional intelligence assessment, At the time, the organizational knowledge of the theory and measurement of emotional intelligence was only beginning to come into existence. Our understanding of psychometric validity and reliability was also limited. In fact, our 'bible' of psychometric validity and reliability was still in the process of being written. The three volume series found in [75], [76], and [77] was not published until 2013. This is not to say that research in this area hadn't been conducted. Rather, many of our customers did not demand the more rigorous approaches outlined in the aforementioned series and elsewhere in the academic literature.

The developer of the original questionnaire provided what they called a validity study. A review of this study many years later found it quite lacking. However, the organization, at the time, accepted the validity study at face value. The psychometric analytic approach employed prior to a full research program being launched at TTI Success Insights in 2017 was adequate enough to see that the performance of the original questionnaire was unacceptable. Several changes were made to the questionnaire in the 2014-2015 time period that provided some marked improvements to its performance.

Admittedly, one of the areas that was lacking in the knowledge base of TTI Success Insights prior to the 2017 time period was an understanding of analysis of covariance and its application to exploratory and confirmatory factor analysis. The limited understanding of the benefits of using such an analytic approach, combined with limited knowledge of emotional intelligence and what it is actually measuring led us to our first technical manual on emotional intelligence and the results found during our first, complete analysis of our EQ model.

Simple issues, such as a minimal number of items used to measure a construct, had largely been ignored. Further, the concept of variance explained was generally a new concept for the organization. Our initial approach was to identify what exactly our questionnaire was measuring and how closely aligned to the Goleman model we actually were. The results of these analyses are documented throughout this work in the relevant historical sections at the end of each appropriate section.

Following this discovery, a process was set in place to update and improve the current questionnaire, as it stood. This means that we recognized, at the time, that we were not fully implementing the Goleman model. We were implementing parts of each of the five main constructs (in our language Self Awareness, Self Regulation, Motivation, Social Awareness, and Social Regulation). However, our questionnaire simply wasn't long enough to have enough relevant questions present to measure all aspects of all things. For example, the 1995 version of Goleman's model lists 25 sub-constructs measuring the five overall constructs. If we assume that a minimum of five items is required for each sub-construct for an adequate measure, then we need a minimum of 125 items and our current questionnaire consists of 57 items.

This last fact is really more a statement about construct representation than it is about construct definition, and we discuss this more in Section 8.3 below. For now, we are presenting the historical *how we got here* story in order to present the argument for why we do, in fact, have solid construct definition evidence of validity. The Goleman model is widely accepted as a valid model for measuring emotional intelligence. As is noted above in Section 2, this model has been defined and re-defined several times over the years as Goleman collected relevant data and evolved his understanding of emotional intelligence. As our understanding is also evolving, we continue to improve our approach to measuring emotional intelligence.

It is our long term goal to continue to develop our EQ model in alignment with the full Goleman model as outlined in Section 3. In particular, we see a longer assessment with items aligned with all sub-constructs in the Goleman model.

8.2 Construct Relevance

The next area to consider in the test content area is that of construct relevance. This is, essentially, the smell-test portion of the validation process. We can consider all the internal structure measures as outlined in Section 11, but those only tell us how well the set of items works together as a whole. This section discusses whether the actual items are, in fact, measuring the construct of interest. The following is from the APA on this topic.

> ... validation focuses on whether each item on the assessment is relevant to the construct tested. [163], pp. 66.

In other words, have we constructed an assessment whose items are actually measuring the intended construct or sub-construct and not just a collection of items that seemingly work well together as measured by, say, an exploratory factor analysis?

The details of each item on the TTI Success Insights EQ questionnaire are obviously proprietary and we do not intend to publish them in this manual. We will present a simple discussion aimed at satisfying the reader that we have constructed relevant items for measuring the various constructs overall. For the sub-constructs, in some cases we must reluctantly admit that we are not directly addressing them and do so in the following section on construct representation.

The TTI Success Insights EQ questionnaire is not designed to trick anyone into revealing anything they are not intending to reveal. We present direct statements in general with a small number of reverse scored items that work well, but really are holdovers from an earlier era and approach. The original intent of the reverse scoring was an attempt to capture less than honest respondents. The idea that reversing a statement and using it to enforce honesty has been debunked over the years. Further, we have established in [29] that intentional introduction of negative language has an impact on the neurological functioning of the brain. It often leads to confusion on the part of the questionnaire respondent.

Current scoring of the questionnaire to generate construct ability scores is currently merely summing the responses, minus one, and then mapping the scores to an interval $[0, 100]$. The forthcoming update to the questionnaire, as described in Section 8.4, comes with the introduction of weighted scoring based on a factor analytic approach. Regardless, there is no trickery and the scoring is straightforward. We literally rely on the garbage in, garbage out model. If the respondent is honest with the questionnaire, the questionnaire effectively provides relevant information to the respondent in the form of a report generated based on the individual's response patterns.

We present the preceding paragraphs to point out that we are relying on the honesty of the individual respondent to provide the relevant information. So, for a sub-construct such as *Accurate Self Assessment* from the Self Awareness construct, we may ask a respondent to rate themself on the item

1. I understand my greatest limitation.

This type question is directly asking whether an individual is willing to honestly assess how well they understand themself. We believe this type item is, in fact, relevant and providing accurate self assessment information. The entirety of the TTI Success Insights questionnaire has been developed in this manner. We seek a set of items that is getting at the heart of the construct or sub-construct definition, following our test construction procedures as outlined in Section 8.4.

8.3 Construct Representation

In this section, we are addressing the concept of construct representation. Directly from the source, see [163], we have

> ... construct representation ... investigates the degree to which the items and tasks on an assessment fully represent the intended construct and do not contain material irrelevant to the construct measured. [163], pp. 66.

We begin with the second part of the definition related to introduction of material irrelevant to the construct being measured. We rely on the following facts:

1. We believe our construct definitions absolutely satisfy the APAs requirements;
2. We believe we have satisfied the APAs requirements for construct relevance;
3. Our test construction procedures are designed with best practices in mind and we follow them rigorously.

The combination of the three above statements lends a great deal of credence to the idea that we have not introduced irrelevant material because our approach to all of the above is designed to avoid such irrelevance being introduced. Our approach is designed to avoid introduction of irrelevant material through the various processes we discuss throughout this document.

The bigger question is related to the first half of the definition. That is, have we covered the entire content domain? The answer to this question is an unfortunate no. As discussed in the section on construct definitions above, TTI Success Insights is not the original developer of this assessment. The original developer unfortunately took a very high level look at EQ as merely the sum total of the five constructs and did not develop the assessment with the sub-constructs at the forefront of their approach.

To put this another way, we were handed a questionnaire and asked to validate it, really going in reverse. Rather than produce the test content validity evidence as part of the overall test construction, we were forced to backfill the arguments. This is an ongoing work in progress. Much success has been achieved in this area and the upcoming version of the TTI Success Insights EQ questionnaire is far superior to that received in the late part of the first decade of this millennium. Unfortunately, at a minimum, the COVID pandemic from 2020-2023 set us back at least that three years, and likely more like five to ten years. We questioned the validity of any data we received during those years and suspended our testing program as a result.

We are now fully back on track and have, as of the writing of this section of this manual, a pilot study scheduled to gather data, test, analyze, and develop scoring algorithms for the next iteration of the TTI Success Insights EQ questionnaire. Our continual improvement process is designed to further monitor and generate continued work toward filling out the entirety of the Golemen EQ model over the next few years.

8.4 Test Construction Procedures

The TTI Success Insights continual improvement program incorporates the test development process as outlined in the diagram presented in Appendix R, as well as following the general guidance presented in this section to help ensure appropriate validity evidence based on test content is developed during the process. In order to properly develop any assessment, one must first understand the underlying theory so that one can adequately define the relevant constructs, see Section 8.1, dictate proper construct relevance, see Section 8.2, and to ensure full construct representation, see Section 8.3. This all begins with an extensive literature review.

Once the literature review is accomplished, one can then go through the process of defining the constructs. Once the constructs are properly defined, it is time to create an extensive item test bank. The creation of the item bank may be a grueling one that requires many iterations before success is attained. It is often advantageous to have a dedicated group of subject matter experts

(SME). The SME group should contain, obviously, experts in the areas covered by the constructs under question. Not so obvious is that the SME group should contain individuals with extensive experience in developing assessments in general, individuals with a great deal of experience in linguistics, logic, and any other area that may be deemed relevant to the process.

As we discuss in Section 8.5 below, the process for identifying an item bank for replacing underperforming items on the TTI Success Insights EQ assessment took several years, in part due to the COVID pandemic, with the number of potential replacement items running well over 100 for the five constructs in question. This is where the next important piece comes, the field test. Once a reasonable set of items have been identified, we test them in our test bank system, also described in Section 8.5 below. In an ideal world, we would have a population all take the same version of the field test assessment. Unfortunately, our current iteration of the test bank system is not designed to handle such a task. We do run small studies by appending test items to the current assessment and gathering data by randomly selecting a predetermined percentage of the TTI Success Insights population.

Once an adequate number of respondent data has been collected, we will turn off the study and analyze the data. Each of the analytic processes we go through are outlined in the relevant sections of this technical manual. We begin with a view of the response distributions to see if we have a reasonable spread across the range of possible responses, a first look at whether the items have adequate item discrimination and difficulty. We further consider statistics such as the α coefficient to assess construct reliability. Factor analysis and item response theory are consulted where appropriate. We also consider inter-item correlations as a confirmation of construct reliability at the item level and item-to-total score correlations as a further view of internal structure validity evidence.

At this point, we take into consideration, in as much as possible, an all encompassing view of the aforementioned information, in conjunction with other information such as the language, grammar, structure, and more, of the items, and make recommendations and decisions on which items to retain and which to remove from consideration. If we do not have enough items to create a full construct or assessment, we iterate the field test with additional items, if any are available, or we go back to the identifying items for the item bank stage and iterate from that stage. This process is iterated until such time as we have a full set of items for a full assessment.

Once the full assessment is available, we then launch a pilot study on the questionnaire as a whole. This consists of gather a larger sample of data in order to accomplish several tasks. The first task is to have a full set of data on the entire assessment based on the same population. The second task is to analyze this data as described above to help establish a baseline internal structure validity evidence argument as well as establishing the internal consistency reliability argument. The third task, assuming the assessment data shows adequate levels of evidence of internal consistency reliability and internal structure validity, is to develop an appropriate scoring algorithm. The scoring algorithm is dictated, generally speaking, by the response format of the assessment. In the case of EQ, we are using a six-point likert-style response format. This dictates that we either go with some form of factor analytic weighted scoring approach, or we could use a more sophisticated graded response model (GRM) scoring from item response theory (IRT).

Given the added complexity of generating GRM based scoring, we have chosen to use a factor analytic approach as outlined in Section 10.7. The complexity is brought into the discussion based on the fact that in order to generate the GRM scoring algorithm, one must first generate the appropriate item parameters via an optimization algorithm, then one must repeat the optimization with the item ability parameter as the focus of the optimization for each respondent. This means

developing the optimization algorithm for use in our internet delivery system (IDS). While such an approach is possible, the potential for errors outweighs the potential benefit of the GRM model given that an equivalent factor analytic based approach is available that requires only standard linear algebraic operations.

At this point, assuming all tasks have been successfully accomplished, we are prepared to launch the assessment. It is also at this stage where we begin the translation and adaptation process as outlined in Appendix Q. This process is essentially what has already been discussed here, iterated in the target language rather than the development language. One item that is not explicitly covered in the diagram in Appendix R is the process of generating equating functions of some form so that scores in various languages may be compared. The topic of equating and linking functions, and their application to adaptation as well as differential item functioning, is discussed in Section **??**.

8.5 Historical Development

The initial TTI Success Insights Emotional Intelligence assessment was developed in the late part of the first decade of the 21st century. It was developed by an independent third party and eventually licensed and purchased by TTI Success Insights around the middle of 2010. As is noted in Section 3, the TTI Success Insights Emotional Intelligence (EQ) model was originally based on the work of Goleman, see, e.g., [81]. Today's version of the TTI Success Insights EQ model is still mostly based on the Goleman 1998 model, although we have attempted to extend the original five constructs into their more granular sub-constructs as we will detail below. It is also noted that while Goleman removed the construct "Motivation" from his model based on the fact that he could not find independent evidence validating its existence as a stand alone construct of EI, our business model has shown this construct to be quite popular in practice. While we acknowledge that it is psychometrically questionable, we do continue to include it in our model for the time being.

The first several years of the EQ model's existence in our business show a slow development of EQ/EI in our marketplace. The performance of the model, as viewed from a statistical perspective, show a relatively unstable model. In 2014, an effort was undertaken to improve the performance of the EQ model by addressing the disconnect that existed among the statements for the various EQ constructs. While some fairly dramatic improvements were seen, the performance was still not at an acceptable psychometric level. We again analyzed the assessment data in 2017 during the development of our first full blown technical manual on the topic of EQ.

It should be noted at this point that the assessment development process at TTI Success Insights was somewhat ad hoc, very much in alignment with the entrepreneurial mindset that had successfully guided the company for its first 30+ years. During this timeframe, a focus was laid on developing a complete continual improvement process that included a redesign of the assessment development and translation (assessment adaptation) processes. Prior to 2017, we essentially followed the majority of our competitors in the industry and reported the alpha (α) coefficient of Cronbach along with some looks at various correlation results. In 2017 and subsequently, we have chosen to align our reliability and validation studies with two main organizations. For our main look at reliability and validity, we have chosen to follow the guidance of the American Psychological Association (APA) as outlined in their three volume series *APA Handbook of Testing and Assessment in Psychology Volumes I, II, and III*, see [75], [76], and [77].

For other areas, such as translation or adaption, differential item functioning, assessment equating and linking, and so on, we have adopted the approach of the International Test Commission (ITC) as outlined in their work *The ITC Guidelines for Translating and Adapting Tests*, see [55], as

well as our synopsis provided in Appendix Q. We have adopted the ITC approach as it is focused much more on the cultural aspects of adaptation versus the language-centric vision provided by a translation based thought process. Further, the concepts of equating and linking are much more prominent in their approach than others, such as the APA which is almost solely focused on English. This is not a critique of the APA approach. It is merely a commentary that multiple sources are usually necessary to cover all one's bases.

We also note that we have outlined our development process in Appendix R. At this point we return to the discussion of the EQ assessment and our attempts to bring it to a psychometrically valid state. As discussed above, we initially encountered the EQ assessment in 2010. We implemented this assessment and worked it into our business over the next few years. A quick analysis of data in 2014 showed a psychometrically underperforming assessment. A series of changes were made to the assessment in early 2015 and we again monitored the performance. In 2017, we pulled the data once again and analyzed it to see that we had some fairly dramatic improvements, yet the performance was still not satisfactory.

At this time, the introduction of our continual improvement process, an updated vision of assessment development, and our decision to align with the APA and ITC highlighted some specific business needs that were not being met. For example, our technology capabilities at the time did not support our continual improvement process, nor did it support our vision of assessment development, item testing, and more. Up to this point in time, if a decision was made that certain statements on one of our assessments needed updating, we collectively decided what the new statements should be and made the changes. There was no testing, external input, third party opinion, etc. In late 2017, the leadership of TTI Success Insights made the decision to support a more robust continual improvement and assessment development process. We developed an internal program that allows the Research and Product Development (R&PD) team to randomly select assessment respondents to respond to test or development items in an effort to gather information on the efficacy of proposed changes without the need to change the assessment while not knowing the changes will, in fact, be successful.

During the construction of our first EQ technical report, the data analysis showed that a subset of 17 of the overall 57 items on the assessment just weren't functioning at a high enough level. At that time, a decision was made to remove those 17 items from the scoring algorithm, leaving them in place and continuing to collect data on those items. Removing these items from the scoring improved the psychometric properties of all EQ constructs to an acceptable level. Acceptable level, in this sense, is meant to convey the fact that all inter-item correlation, item-to-total correlation, α-coefficient scores, and factor loading and variance explained statistics were at psychometrically valid levels. We then published the technical report and ran the EQ assessment scoring on the 40 psychometrically valid and reliable items.

During the first half of 2018, the TTI Success Insights Technology team worked with the R&PD team to design and implement our *Test Bank* system. Once the system was deployed and operational, the R&PD team began a series of tests aimed at improving our EQ assessment, among many other goals. The initial efforts were slow with, at best, moderate levels of success. Most of 2019 was spent attempting various studies to determine if we could, in fact, improve the current constructs without the need to start completely over. Towards the end of 2019 and into the beginning of 2020, we began to see some dramatic improvements in our efforts at properly defining constructs and then developing statements that got to the heart of the constructs. This is really getting at the core of the discussions around Sections 8.1 through 8.3 above. Are the constructs properly defined? Are statements we are developing relevant and representative of the aforementioned constructs?

Unfortunately, we were to find out that we would need to wait for some time to actually answer these questions. Just as we were beginning to see some very successful results, the COVID pandemic ground us to a halt, both proverbially and in reality. Proverbially, we were concerned about the emotional state of mind of a population being forced into seclusion by an ineffective, world-wide, governmental response. Anecdotal evidence of an emotionally unstable population abounded. Reports of dramatic increases in substance abuse and domestic violence were commonplace. We decided as an organization to suspend our testing phase for all our assessment improvement projects until such time as we could determine whether and to what extent any of our data was being impacted by the pandemic.

Additionally, our business ground to nearly a halt. So, the reality was that even had we decided to continue our testing processes, the velocity of EQ respondents would not have been able to support a testing phase in anything resembling a reasonable time period. In early 2023, we conducted some ad hoc data analysis to see if the pandemic had any major impact on our data. The results have been somewhat mixed. We have seen some anomalies in specific small groups such as teachers, students (university level), first responders, and medical professionals who were directly in contact with the victims of the pandemic. However, in the aggregate, we did not see a major impact in the overall EQ data analysis. We began to reinstitute our testing program and began to go full ahead with improving our EQ assessment items.

During the period of 2019-2023, we tested several hundred EQ statements in various forms. In many cases, original statements suffered from poor language usage or grammar. For example, several statements on the 2015 version of our assessment were *negatively keyed*. One of our first attempts focused solely on determining to what level this negatively keying of statements impacts the psychometric performance of a particular statement. Many of these results are documented in the article [29] and, in part, show the confusion that can arise when stating something in reverse that should otherwise be relatively straightforward. The original decision to use this idea of negatively keying an item is from an archaic approach to identifying individuals who are being less than honest in their assessment responses.

We also encountered poorly phrased statements, statements with double negative structures, ambiguous statements, and more. Further, we found that as we gathered data, we were having a great deal of success well beyond replacing the 17 items mentioned above. We simplified language, removed negatively keyed items in as much as possible, improved the grammar and usage, and attempted to identify any construct irrelevant statements and remove them. The final tally is to make changes to 34 of the current 57 items. There were originally 11 negatively keyed items and we have removed 4 of those. Further, we attempted to get closer to the core constructs and sub-constructs as identified by Goleman in Table 2.1 above.

We will not detail the specifics of the changes to individual items as we view that as proprietary. Further, any of the psychometric results are presented in the relevant sections on evidence of reliability, see Section 6, evidence of internal structure validity, see Section 11, and evidence of relationships to external variables validity, see Section 12.

9 Item Response Theory

The purpose of this chapter is to outline the various item response theory approaches that are applied to the TTI Success Insights suite of assessments. Our suite of assessments may be thought of in two very general categories, in some sense. These categories are polytomous response models and forced-rank response models. The focus of this chapter is on the former, with the latter reserved for a discussion at a later time.

As noted earlier in this manual, the TTI Success Insights Emotional Intelligence (EQ) assessment is a questionnaire using a 6-point Likert-style response format. Such a response format may not be modeled, strictly speaking, by a polytomous response model. During the process of testing and updating the EQ technical manual, a host of data is to be collected and analyzed in a Likert-style format. It is the intent of this process to use the tools available to the suite of polytomous response models to analyze the EQ response data as part of the decision making process behind building the next generation of this assessment.

Prior to a discussion of the specifics of the TTI Success Insights approach to polytomous response models, it is instructive to discuss some of the motivation and history behind Classical Test Theory (CTT) and Modern Test Theory (MTT), the latter also known as Item Response Theory (IRT). In the follow sections, we generally follow the works in [10], [129], and [132].

The article in [132] is a synopsis of polytomous IRT models and is part of a larger series on quantitative applications in the social sciences. The focus of [129] is a compilation of knowledge of polytomous IRT models, the various tracks of development, and suggested approaches to testing data and model fit. The discussion in [10] centers on the computational techniques used to estimate the parameters of the various IRT models, their characteristic and information functions, as well as scoring algorithms.

It is not the intent of the authors of this manuscript to provide a complete and comprehensive history of the development of IRT, nor is it the intent of the authors of this manuscript to provide a complete synopsis of the current state of IRT models. For this, we refer the reader to the aforementioned works. In the following, we provide references, where appropriate, as stated in the original works as applicable.

9.1 From Classical Test Theory to Item Response Theory

Some authors attribute the foundation of CTT to Cattel around 1890, see [41] and [132], while others trace it to the work of Spearman in the early part of the 20^{th} century, see [164]. Regardless of the exact origin, the underlying problem is essentially the same: a desire to formalize measurement theory for mental tests. The primary issue that CTT attempted to understand was the measurement of error associated with mental test scores.

CTT hypothesized that the existence of errors could be stated in the following relation

$$X = T + E, \tag{9.1}$$

where X is an individual's observed score, i.e., something we can measure such as a test score, T is the individual's (unobservable) true score, and E is the error associated with the true score. With these definitions, we may then define the classical version of test reliability by denoting the variance in the true and observed scores as σ_T^2 and σ_X^2 and defining

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \tag{9.2}$$

which may, in some sense, be interpreted as a signal to noise ratio.

One of the main issues with CTT and its application in more contemporary assessment/data analysis is that CTT often over-simplifies a given situation in its approach. For example, many classical test theory approaches rely on some form of dichotomization during the process of analyzing. Take, for instance, the classical definition of item difficulty. Item difficulty is often defined simply as the percentage of incorrect responses to an item by the population under consideration. This is a straightforward definition if the test item of interest is simply a yes/no or right/wrong or true/false type question. However, more contemporary formulations rely on what may be termed graded responses or responses on a Likert-type scale.

The idea of "dichotomizing" a variable is a simple one. However, it should be, at least, intuitively clear that something must be lost during the process. As an example, take an assessment item that is originally scored on a Likert-type scale with six possible graded choices. The first "bucket" is denoted by 1 and represents the lowest level of agreement with the item, with each successive bucket representing a slightly higher level of agreement until the final bucket is reached representing the most agreement with the item.

From the CTT standpoint, we are generally forced to choose a dichotomization, usually by defining any items ranked 1, 2, or 3 to denote a negative response, and items ranked 4, 5, or 6 to be positively endorsed. In this way, we have created a binary variable, 0 indicating generally unfavorable response, and 1 indicating generally positive response. We provide a simple example to illustrate the potential pitfalls and to motivate the remaining discussion.

As mentioned above, item difficulty may be defined simply as the percentage of negative responses to a (dichotomous) item, the thought being the higher the percentage, the more difficult the item. It is noted that this is not the only definition of item difficulty, nor is it the most appropriate. However, it is the simplest and most useful for a simple explanation of the potential issues that accompany dichotomization of a variable. Suppose we consider two separate response distributions, both with 100 respondents, on a six-point Likert-type assessment item. Suppose the response frequencies are given in Table 9.1.

Table 9.1: Response Frequencies
Dichotomization Example

Subset	1	2	3	4	5	6
X	10	15	25	25	15	10
Y	35	10	5	5	10	35

A simple transformation takes the information from Table 9.1 and creates Table 9.2. Simply summing the values in Table 9.2 for categories 1,2, and 3 provide on estimate of item difficulty for the two distributions. A quick check shows that both the X and Y distributions show a difficulty of 0.50.

TTI SUCCESS INSIGHTS®

Table 9.2: Response Distribution
Dichotomization Example

Subset	1	2	3	4	5	6
X	0.10	0.15	0.25	0.25	0.15	0.10
Y	0.35	0.10	0.05	0.05	0.10	0.35

However, it should be clear that the behavior of these two items is quite different in nature. The first (X) is a distribution similar in nature to a normal distribution, while the second (Y) is a barbell shaped distribution with the extremities dominating. This latter example is an example of an all or nothing response pattern, either one loves the item or one hates it. The former exhibits, in some sense, an ideal distribution with a majority of the distribution centered around the middle of the response scale. Regardless, information appears to be lost through the process of dichotomizing a variable.

Given the above described potential problem, it is the desire of item response theory (IRT) to extend the ideas of CTT in a reasonable manner while providing a framework that both accounts for classical test theory results where applicable, yet allows for the extension of these ideas to a modern approach to the classical analysis. According to [132], IRT may be described as relating the probability of a person responding to an item in a specific manner to the standing of that person on the trait the item is measuring. This is a big jump on the theoretical side from attempting to measure the error in true score, based on the original definition of an observed test score, to relating a respondent's probability of answering in a specific way to their "ability" level as measured by the item.

Reading through [94] and [132], one gets a feel that there are slightly differing opinions of the evolution of IRT over the years. For example, [132] traces the foundation of IRT to Fechner circa 1860 (see [9]) and Thurstone in 1925 (see [18], [170], and [176]). The mathematical roots of IRT in psychology may be traced back to the work of Binet, Simon, and Terman from 1916 (see [9] and [18]), while the formal basis of IRT is generally attributed to Lawley for his work in 1943 (see [9], [107], and [176]). Lord (see [112]) formalized IRT as the extension of CTT (see [9]). Lord ([113]), Lord & Novick ([114]), and Birnbaum ([17]) helped establish the acceptance of IRT in the psychological community. Finally, Rasch (see [141], [142], and [143]) provided the first specific class of IRT models.

Compare the presentation of [94], which discusses more of the contemporary development of IRT beginning with Lord (see [110], [111], and [112]), citing that Lord led the transition from CTT to IRT from Gulliksen's (see [93]) challenge to develop invariant item statistics for test development. [94] also notes the development of the "item characteristic curve" (see [174]) and the use of the term "latent trait" (see [108]). Finally, a suggestion to change the terms "item characteristic curve" and "trait" to "item response function" and "ability" is provided by Lord (see [113]).

Note that it is not the intention of the authors of this chapter to suggest that there are any inconsistencies in the telling of the storing of the development of IRT. Rather, it is to point out the differing perspectives and to suggest both as important reading for the individual interested in understanding the role and development of the theory. On one hand, [132] provides a somewhat more motivation based approach, while on the other, [94] seems to present a more milestone based perspective. Both are equally important in tracing out the history.

The majority of the presentation is still related to the mostly dichotomous approach to item analysis. Even at this stage, we still define concepts such as difficulty in terms of a binary approach, we just



now compute an item characteristic curve or item response function, depending on one's preference for the language used. However, it is this transition from CTT to IRT which marks the beginning of the ability to take the leap from a dichotomously based analysis to polytomous response based analysis.

9.2 A Short Overview of the Rasch Models

As with much of the earlier section, the early development of IRT models centered around binary variables or dichotomously scored items. There is a right and a wrong answer, or the data may be aggregated in such a way as to be interpretable in such a manner. In some sense, this is rather convenient for analysis. As mentioned in an earlier example, one very simple definition of difficulty is to simply take the ratio of incorrect to total responses as an approximation.

In this way, we are equating difficulty with the probability of an incorrect response to an item. In the setting of a rating scale or Likert-style responses, one may formulate the same definition of difficulty by defining correct (incorrect) as the ratio of positive (negative) endorsement to total responses. This does cause some issues in the cases where the number of response categories is odd. For sake of simplicity of exposition, we assume an even number of response categories noting that alternative definitions for item difficulty do exist that work well with both even and odd number of response categories.

In the previous section, we presented the idea that IRT may be thought of as relating the probability of a person responding to an item in a specific manner to the standing of that person on the trait the item is measuring. If we assume that this may be represented by the cumulative distribution function (CDF) of a probability distribution, such as a normal or Gaussian distribution, we may plot along the x-axis the ability measured by the latent trait and along the y-axis the CDF, see Figure 9.1.

There are some interesting characteristics of a CDF that have natural interpretations in item analysis. Note first that the CDF displayed in Figure 9.1 is defined on all $\mathbb{R}$. Not all continuous probability distributions are defined on the entire real line. However, the two most prevalent distributions in IRT literature, the logit and probit distributions, are defined on $\mathbb{R}$.

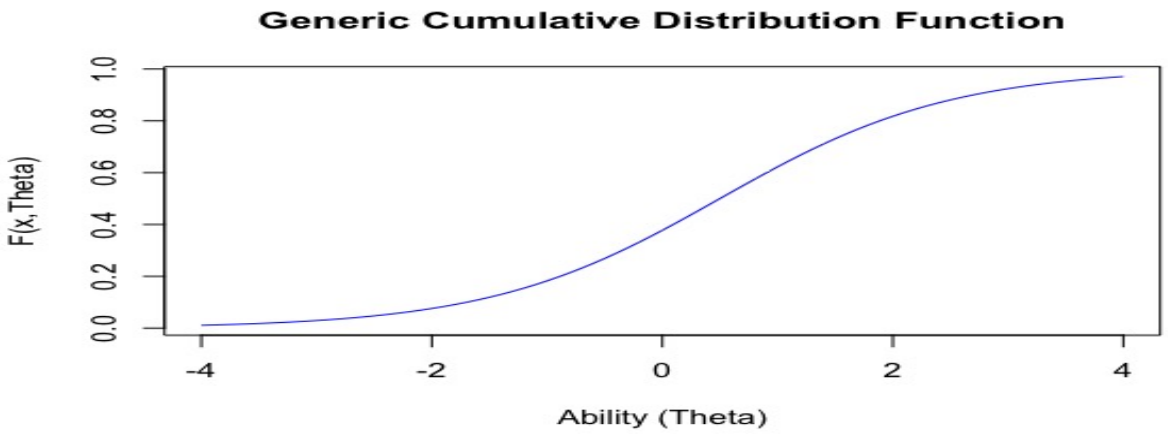


Figure 9.1: Generic CDF

TTI SUCCESS INSIGHTS®

Note second that there is a point, call it x_0 in $\mathbb{R}$ such that the distribution is concave up (the graph "opens up") on $(-\infty, x_0)$ and concave down (the graph "opens down") on (x_0, ∞). The point $(x_0, F(x_0))$ is known as an inflection point for the function $F(x)$. At this point, some interesting things happen for a generic CDF.

$F(x_0)$ is the half way point between the minimum and maximum values of $F(x)$ and, assuming the function $F(x)$ has a derivative $F'(x)$, $F'(x)$ attains its global maximum at (x_0). The maximum is, in fact, global given that $F(x)$ is strictly monotonically increasing on $\mathbb{R}$.

With the interpretation of IRT as relating probability to ability, we an infer for a generic CDF that the range of $F(x)$ is $(0, 1)$, which implies that $F(x_0) = 0.50$. In other settings, the CDF may have various parameters that may come into play determining a probability other than 0.50 for this relationship. For example, for the Rasch 3-PL model which is introduced later in this section,

$$F(x_0) = \frac{1+c}{2},$$

where c is a parameter in the 3-PL model representing a lower bound. This lower bound represents the possibility of guessing, say in a multiple choice assessment format. Note that in our format, we have implicitly assumed no guessing implying a value of $c = 0$.

This special point, the point x_0 defining the inflection point of the CDF $F(x)$, is known as the item location. One main advantage of this formulation is that the item location and the individual ability are now measured on the same scale or metric. This means that we may use the previous interpretation of IRT, that of relating probability to ability, to now state that if an individual's ability is greater than the item location, the individual is more likely than not to respond "correctly" or in a "positive" manner to the item. In other words, we have extended the definition of item location or item difficulty from the CTT viewpoint to a modern version in the IRT framework.

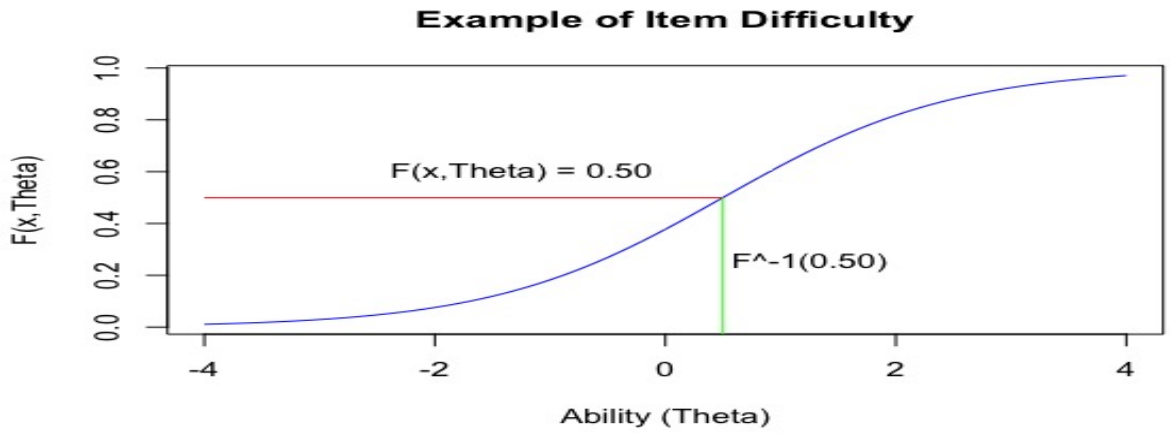


Figure 9.2: Example of Item Difficulty

Figure 9.2 is a continuation of Figure 9.1 with the vertical green line representing the point on the x-axis x_0 and the horizontal red line representing the value on the y-axis $F(x_0)$. The reality is that we will not know x_0, but will know the value for $F(x_0)$. In this case, we need to find the value of x corresponding to the equation $F(x) = 0.50$. Since $F(x)$ is monotonically increasing, F is invertible, and we may find $x = F^{-1}(0.50)$ to be the value of x_0. More generically, we find the solution to the

TTI SUCCESS INSIGHTS®

equation $F(x) = y$, where y is the midpoint between the lower and upper asymptotes of the CDF. Given that the CDF is strictly monotonically increasing (see Definition 9.1 below), this solution always exists.

Another notion of interest in item analysis and IRT is that of item discrimination. Item discrimination is a statistic that measures how well an item is able to differentiate individuals along the trait continuum. In other words, given a small change in the input (i.e., the trait), how significant or large is the change in the probability of giving a correct response or positive endorsement? The larger the change in output given the same change in input, the more discriminating an item is.

The preceding description of item discrimination foreshadows one possible approach to computing it. A natural computation to measure the changes previously discussed is to take the ratio of the change in output to the change in input, or the slope of a line. In particular, we are interested in the slopes of the lines tangent to the CDF. Moreover, we are interested in where this infinitely large set of lines attains maximum slope.

As noted above, the derivative of the CDF attains a global maximum at the point x_0. Given the usual interpretation of the slope of the line tangent to a curve being equal to the value of the derivative of the function generating the curve at the point of tangency, we have our suspect for the value of item discrimination. The value of item discrimination is, in fact, defined to be the slope of the line tangent to the CDF curve at the point where the derivative of the CDF attains a global maximum, x_0. See Figure 9.3 for a visualization of the preceding discussion.

Similar to finding the value for which $F(x) = 0.50$, at least in this example, we now find our item discrimination value to be the value of the slope of the line tangent to the CDF at x_0, or $m = F'(F^{-1}(x_0))$. For completeness of exposition and to provide the details for the reader unfamiliar with the some of the mathematical constructs discussed above, we finish this portion of the discussion with a presentation of the following theorems without proof.

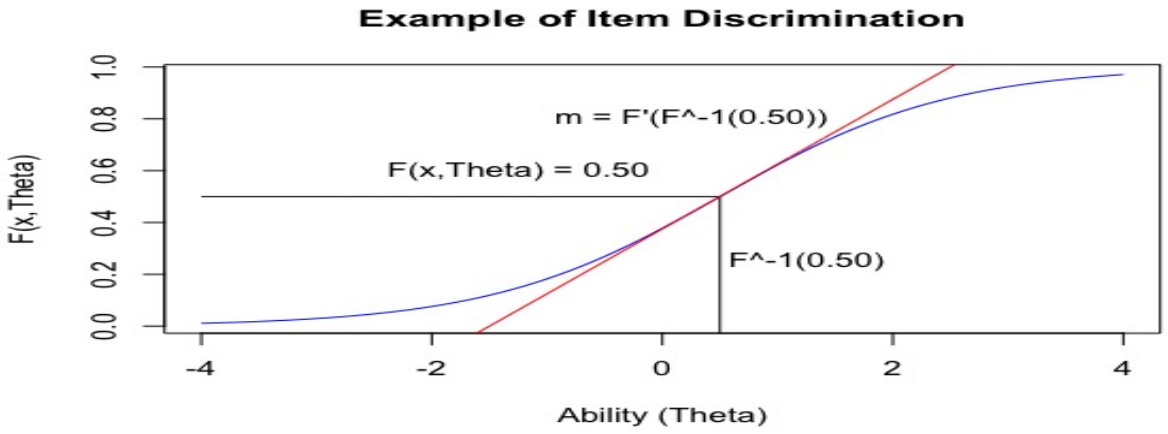


Figure 9.3: Example of Item Discrimination

Definition 9.1. *Let $f : \mathbb{R} \mapsto \mathbb{R}$. The function f is said to be (strictly) monotonically increasing if for all $x \leq y$ $(x < y)$, we have $f(x) \leq f(y)$ $(f(x) < f(y))$.*

We state without proof the following theorems courtesy of J. Quigg. One may also consult any of

TTI SUCCESS INSIGHTS®

the multitude of available advanced calculus or introductory real analysis text books for a discussion of these and other similar topics.

Theorem 9.2. *Let I be an interval and let $f : I \mapsto \mathbb{R}$. If f is monotone and $f(I)$ (the image of I under f) is an interval, then f is continuous.*

So, assuming that a function is monotonic and maps an interval to an interval, then the function is continuous.

Theorem 9.3. *Every continuous 1-1 real-valued function on an interval is strictly monotone.*

Theorem 9.4. *Every continuous 1-1 real-valued function on an interval has a continuous inverse.*

It is this third theorem that we need to invoke for some of the previous statements. It is the combination of the first and second theorems that provide for one possible approach to the proof of the third.

Now that we have provided some of the motivation behind the development, we provide some concrete examples in the dichotomous space prior to extending these concepts to the polytomous space. We begin by noting that one of the first researchers to provide a specific class of IRT models for the dichotomous case is Rasch in his 1960 work, see [141]. There are many equivalent ways to write down this model, so we choose the following notation:

$$Pr\{X_{ni} = 1\} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \tag{9.3}$$

where θ_n represents the ability level of the n^{th} respondent and δ_i represents the location or difficulty of the i^{th} item of an assessment. The functional relationship represents the probability that the n^{th} individual chooses a "correct" answer on the i^{th} item and is denoted by $Pr\{X_{ni} = 1\}$, but may be more explicitly denoted $Pr\{X_{ni} = 1|\theta_n\}$. The notation $Pr\{X_{ni} = 1|\theta_n\}$ makes explicit the conditional probability's dependence on respondent ability.

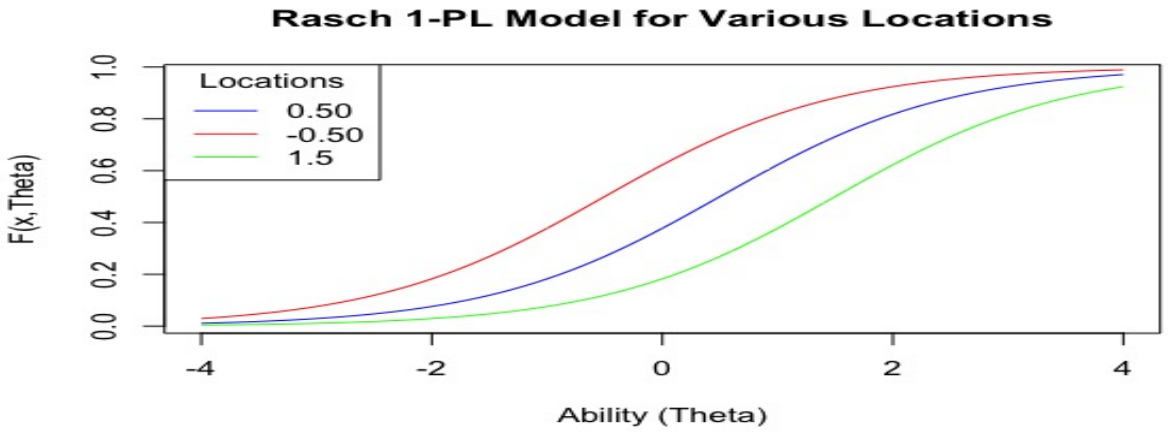


Figure 9.4: Rasch 1-PL Model

Figure 9.4 presents the Rasch 1-PL (one parameter logistic) model for three separate values of the location or difficulty parameter. The alert reader will have noted that the blue curve in the figure is

TTI SUCCESS INSIGHTS®

identical to the blue curve presented in the earlier examples. The red curve is "easier" in the sense that it requires a lower level of the assumed underlying latent trait to obtain the probability of 0.50 of obtaining the "correct" answer to the given item. The green curve is correspondingly harder for similar reasons. Note the implicit assumption in the Rasch 1-PL model the item discrimination is fixed and equal to one.

Changing notation slightly, we introduce the concept of a discrimination parameter and present the Rasch 2-PL model in the following form.

$$f(x) = \frac{\exp\left(a\left(x-\delta\right)\right)}{1+\exp\left(a\left(x-\delta\right)\right)}, \tag{9.4}$$

where a now represents the discrimination parameter.



Figure 9.5: Rasch 2-PL Model

Figure 9.5 displays the Rasch 2-PL model for various values of the discrimination parameter a. The $a = 1$ value corresponds to the 1-PL model. The green curve represents a more discriminating response function, while the red curve shows a not very discriminating response function.

A final example is the Rasch 3-PL model which incorporates a "guessing" parameter that provides for a lower asymptote other than zero. The usual example for this is when an individual is taking a multiple choice assessment in which each item has four possible answers. In this case, one would expect that the lower limiting behavior to be 0.25. In this case, we modify the 2-PL model from (9.4) as follows:

$$f(x) = c + \frac{(1-c)\exp\left(a\left(x-\delta\right)\right)}{1+\exp\left(a\left(x-\delta\right)\right)}. \tag{9.5}$$

It is instructive to spend a moment exploring the formula in (9.5). Note that if we set $c = 0$ we recover exactly the formula in (9.4) for the 2-PL model. Similarly, if we set $c = 0$ and $a = 1$ in (9.5), we recover the formula in (9.3) for the 1-PL model. Figure 9.6 shows an example of the 3-PL model with $c = 0.25$.

There is an interesting and important pattern emerging here. We began with yes/no type questions and defined certain statistics such as item difficulty and discrimination. These were only defined on dichotomously scored items and provided limited information. We generalize in such a manner

TTI SUCCESS INSIGHTS®

that the interpretation of the model is slightly different, yet we still arrive at those very important statistics, in some sense. We then generalize a bit more and still arrive at models, that when simplified, are completely compatible with the previous versions.

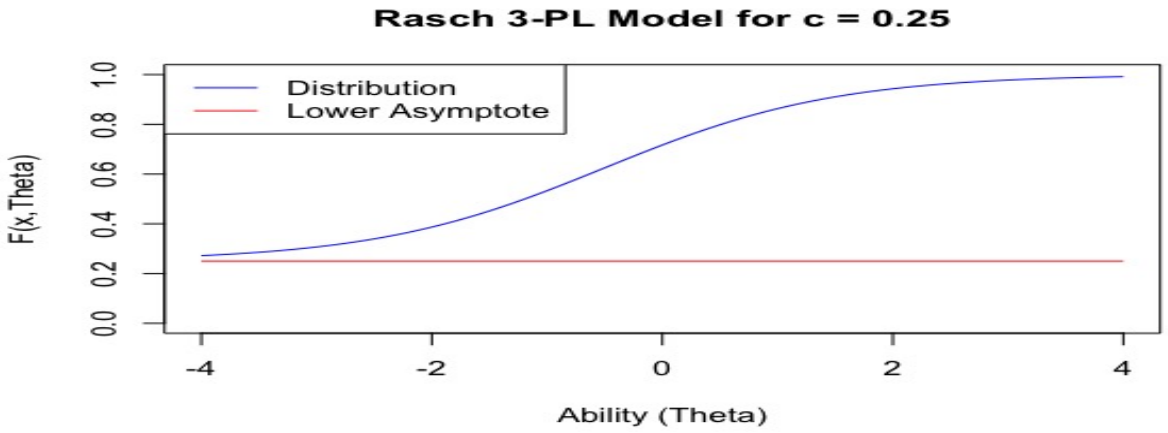


Figure 9.6: Rasch 3-PL Model

Note that there are other possible choices for the distribution function under consideration. The Rasch series of models typically uses the logistic function, hence the descriptive names n-PL model which is short for the n-parameter logistic model. In practice, there are only the 1, 2, and 3-PL models. However, at least theoretically although not much used in practice, is the 4-PL model. The 4-PL model incorporates an upper asymptote that is less than one. However, this model has not gained much traction, at least not that the authors of this chapter are aware.

The Rasch models increase in complexity by considering varying numbers of parameters with various roles. At each stage, the earlier models are recoverable by judicious choice of parameter values, which is a very appealing feature. However, the underlying assumption is that we must, in order to apply the Rasch family of models to our problems, still dichotomize our data. We require a solution that allows us to not lose the valuable information that is being presented to us in the ordered category response assessment data we collect. This is where Samejima and her myriad contributions to the field begin to emerge.

9.3 The Transition from Dichotomous to Free Response Models

Before diving directly into the polytomous models of Samejima's approach, it is instructive to consider the underlying and fundamental differences in motivation behind the construction of the various models. According to [132], pp. 62,

> ... the Rasch models have been built on the requirement of model parameter separability, with the associated issues of sufficiency, specific objectivity, and fundamental measurement. Any hypothesized response process that may have generated the data to be modeled is essentially an afterthought.

In this setting, the term "specific objectivity" refers to the assumption that the underlying trait being measured should be independent of any particular instrument through which it is measured.

TTI SUCCESS INSIGHTS®

It is interesting, in light of contemporary views on the importance of response processes, that Rasch developed his theory with "any hypothesized response process" being an afterthought. In comparison, Samejima's approach is quite different and is summarized in the following statement, also from [132], pp. 63:

> In Samejima's framework, however, the plausibility of the response process thought to generate the data is paramount, because for Samejima ([157]), the main role of a mathematical model in psychology is to plausibly represent psychological reality. Thus, the hypothetical response process is not ancillary speculation as it tends to be for the Rasch models; rather, it is the foundation for her whole approach to modeling polytomous data.

So, to provide a short summary, the response processes are an afterthought in the formulation of the Rasch models, while they are paramount in Samejima's framework. This is not to say that the Rasch models have no use. In fact, as the reader shall see, Samejima essentially uses the Rasch model framework to model the dichotomous boundary characteristic functions used to build the polytomous item response theory models of her theory. It is interesting, however, to note the fundamental difference in underlying assumptions that led to the development of quite similar (mathematical) models in the end.

In much of the literature regarding the polytomous item response theory models, little mention is made to specific models or specific response formats. The reader familiar with TTI Success Insights assessments that utilize an ordered response model or Likert-type model should keep this response format in mind while reading the remainder of this section. For the reader unfamiliar, TTI Success Insights typically uses a format intended to measure level of agreement with a given statement. For example, in measuring one's agreement with how important the statement "Bold, daring" is related to one's behavior, we might ask a respondent to rate on a scale of 1 to 6 with 1 being anchored with the statement "Not at all like me" and 6 the statement "Very much like me".

During the forthcoming discussion, the authors of this chapter suggest that the reader keep the preceding example in mind. As a foreshadowing, from this example, we would be considering six categories with five boundary or decision points for the respondent.

There are two points of interest to make prior to presenting the IRT models. First is that we desire and have the ability to collect data in more than a yes/no or right/wrong format, for example on a Likert-type response scale. Moreover, we wish to analyze this data in a way that does not lose information. As demonstrated in an earlier example, much more information is contained in a six category format versus a dichotomous response.

Further, Samejima (see [156], [157]) demonstrates the increase in information available when gathering the information with more categories (polytomous), while Cohen (see [46]) shows that there exists a systematic loss of information when dichotomizing continuous or categorical data with many categories. With these concepts in mind, we present a condensed version of Samejima's graded response model theory.

Generally speaking, each response category of an item exerts a level of attraction on each individual respondent. The level of attraction varies across the population of respondents based on their individual level of the underlying trait being measured. An individual's level of attraction in a given category is contingent upon the respondent having been attracted to the previous category. In other words, one is not immediately attracted to the level six category in our example without first having considered and found each previous category attractive enough to at least consider the next higher level.

Samejima defines what she called a processing function, denoted $M_{ij}(\theta)$ which is defined to be the probability of being attracted to category j over and above category $j-1$. It is helpful to make note that historically, Samejima used the notation $M_{ig}(\theta)$. Given that the primary author of this manuscript is trained in mathematics, a decision is made to rewrite many of Samejima's formulae in a manner more conducive to the mathematical structures the authors are familiar with. Historically in mathematics, letters such as $f, g, \& h$ are reserved for functions, while letters $i, j, \& k$ are typically reserved for dummy indices. While g in $M_{ig}(\theta)$ is not a dummy index, later in this work there will be need to use dummy indices for summations.

In the original works of Samejima and in the reproduction, e.g., in [132], the authors have a tendency to reproduce exactly the equations as originally written. This is not intended to be a criticism of the original work of Samejima, nor of anyone reproducing her work as in [132]. It is often the case that original notation is presented as part of preservation of historical work and as a testament to the importance of the work. The authors of the current manuscript find it easier to read and comprehend by using more historically consistent notation from the standpoint of mathematics.

The overarching idea is that the respondent is successively accepting and rejecting response categories based on the idea of the processing function. In the end, there is a cumulative effect that must be accounted for in order to determine the true probability of an individual of a given trait level to ultimately select category j over and above category $j-1$, but to reject continuation to category $j+1$. According to [132], pp. 64,

> The cumulative attraction is operationalized as $P^*_{i_j}(\theta)$, where $P^*_{i_j}$ is defined as the probability of responding positively at a category boundary given all previous categories, conditional on θ. ... the probability of responding in a category can then simply be calculated as the probability of responding positively at a category boundary minus the probability of responding positively at the next category boundary...

In terms of the $P^*_{i_j}$, one may write this as

$$P_{i_j} = P^*_{i_j} - P^*_{i_{j+1}}, \tag{9.6}$$

where P_{i_j} is to be interpreted as the probability of a respondent of a given trait level responding in category j. In this sense, P_{i_j} is logically modeled "... as the difference between the cumulative probabilities of serial attraction to two adjacent categories."

There are two broad classes of models in the Samejima framework, the homogeneous case and the heterogeneous case. This work largely discusses the homogeneous case with a brief synopsis of the heterogeneous case to finish this section. It is necessary to make note of the confusion in the literature surrounding the use of the term *graded response model* or GRM. In her original work, GRM refers to the class of models overall.

In fact, the editors of [129] state, "In effect, the graded response model is, for her, a model of models", on pp. 5 of their introduction to the collection. In more contemporary usage, the term GRM is used to denote, specifically, Samejima's homogeneous logistic model. So, in the literature, GRM refers to a specific case of Samejima's overall vision of the polytomous IRT model, while in Samejima's work the same term refers to the general framework she developed.

There are two specific characteristics of the overall framework. First, the homogeneous models are the only models operationalized by (9.6). Second is the fact that all homogeneous models have as their foundation the functions $P^*_{i_j}$, which all have the same shape and hence the choice of the descriptor "homogeneous". As noted in the previous paragraph, GRM in the wider body of



literature on this topic refers to Samejima's homogeneous logistic model, and we shall follow that convention.

Prior to presenting the mathematical particulars for this model, it is instructive to summarize some of the history as presented in [132]. GRM is the archetypical model in the Samejima framework due to the following:

1. GRM is one of Samejima's original operationalizations of her theory, [154];
2. GRM is the polytomous manifestation of Thurstone's method of successive intervals, see [38], [65], and [147];
3. GRM relies on (Rasch) 2-PL model, see e.g., [56], [71], [72], and [159], to obtain cumulative boundary functions P^*_{ij};
4. The functions P^*_{ij} could be built based on any appropriate mathematical function, see [155], [158];
5. In practice, only 2PL and 2-parameter normal ogive have been used; both outlined in Samejima's original work, see [154].

Following the lead of the 2-PL model given in (9.4) we define the category boundary response function (CBRF) of the GRM to be

$$P^*_{ij}(x) = \frac{\exp\left(a_i\left(x - \beta_{ij}\right)\right)}{1 + \exp\left(a_i\left(x - \beta_{ij}\right)\right)}, \tag{9.7}$$

where the a_i represent the discrimination parameter of the i^{th} item and β_{ij} denotes the location of the j^{th} boundary for the i^{th} item.

With the CBRF at hand, we can now go after the true desired quantity, the item characteristic response function (ICRF). Recall that the ICRF measures the true probability of responding in a given category as compared to the probability of crossing the boundary as in the CBRF. In order to do so and utilize the definition provided in (9.6) we must first define two quantities of interest.

In order to compute P_{i_1}, we must compute $P^*_{i_0} - P^*_{i_1}$, yet we have no definition for $P^*_{i_0}$. One may think of $P^*_{i_0}$ as the probability that a individual responds to the item in the lower category or higher. Given that a respondent is, by definition, responding to the item in the lowest category or higher, we may simply define $P^*_{i_0} = 1$ and define its location to be $-\infty$.

In a similar manner, in order to define P_{i_n}, where n is the total number of boundaries, we need to compute $P^*_{i_n} - P^*_{i_{n+1}}$ and we have not defined $P^*_{i_{n+1}}$. However, the interpretation of $P^*_{i_{n+1}}$ is for a respondent to answer the item in a category higher than the highest available category. Hence, we define $P^*_{i_{n+1}} = 0$ as the interpretation is only possible if this is the case. Also note that the location for $P^*_{i_{n+1}}$ is ∞.

Thus, we have the following ICRF for the GRM model in general, and specifically for the case of P^*_{ij} as provided for in (9.7). $P_{i_1} = P^*_{i_0} - P^*_{i_1} = 1 - P^*_{i_1}$, $P_{i_n} = P^*_{i_n} - P^*_{i_{n+1}} = P^*_{i_n}$, and the remainder are as defined in (9.6). To summarize,

$$P_{i_j} = \begin{cases} 1 - P^*_{i_1} & \text{if } \ j = 0, \\ P^*_{i_n} & \text{if } \ j = n, \\ P^*_{i_j} - P^*_{i_{j+1}} & \text{otherwise.} \end{cases} \tag{9.8}$$

TTI SUCCESS INSIGHTS®

There are a handful of convenient properties to mention at this point. The fact that the model has only a single discrimination parameter per item implies that the shape of all the CBRF is the same and it is the location for each CBRF that varies. This is convenient in that the location for each CBRF for the GRM model is solved in the same manner as previously. Since each CBRF is a dichotomous model, and we piece together the cumulative probabilities to generate the ICRF, it is tempting to look at the ICRF for the location. In fact, one should look no further than solving the equation $P^*_{ij} = 0.50$ as described in an earlier section.

In contrast, the polytomous Rasch models have boundary locations where adjacent ICRF cross. One nice result of this GRM property is that there are never boundary reversals, which may occur in other models. Another nice property is that GRM models may have different discrimination parameters for different items on the assessment in addition to multiple measures of location (difficulty).

As a final note, it is possible to consider a free response or continuous response model. This can be accomplished by a limiting process in that we may consider the limit as the number of categories approaches infinity. In this way, perhaps a respondent answers on a sliding scale with endpoints 0 and 10 (or 0 and 1, or 0 and 100). Rather than choosing an integer between 1 and 10, one simply slides the ruler to an appropriate place on the continuum that is consistent with one's viewpoint on a given item. So, rather than answering 2 out of 10 on a categorical scale, one may slide to 2.10304657 out of 10. While Samejima explored this approach, it has not caught on as a popular one in practice. This topic is beyond the scope of the current discussion, but may be the topic of a future paper by the current authors.

9.4 Item and Test Information

One key area that assessments are required to address is, as noted in Section 9.1, the concept of reliability. Typically, one uses a single metric such as the α coefficient of Cronbach, see [58]. More advanced treatments relying on a factor analytic approach may consider McDonald's ω, see [179]. These are concepts in the classical test theory (CTT) space. A major contribution of IRT is an extension of this concept.

IRT introduces item information (IIF) and test information functions (TIF) as generalizations of the reliability argument from an average across the entire assessment to gathering information across the spectrum of the latent trait on an item-by-item basis. Related concepts may also be introduced such as item characteristic curves, item category characteristic curves, boundary characteristic curves, and more.

The types of curves or functions one may be interested in discussing is related to the type of IRT model employed. For example, in the dichotomous case we have that the item information functions may be easily written down or may be derived from the probability models discussed in Section 9.1. Recall the amended notation from that section and let $P_i(\theta)$ denote the probability model for the 1-PL model as shown in (9.3). Let $Q_i(\theta) = 1 - P_i(\theta)$. According to Fisher information theory [70], the information function for the 1-PL model is given by

$$\mathcal{I}_i(\theta) = P_i(\theta)Q_i(\theta). \tag{9.9}$$

Note that the item information function in this case is related to the item characteristic functions defined by $P_i(\theta)$. [10] gives an excellent introduction from the viewpoint of item characteristic functions. In particular, one can show that for the 1-PL model

$$\mathcal{I}_i(\theta) = \frac{\partial P_i(\theta)}{\partial \theta}. \tag{9.10}$$

Each item on the assessment has an item information function. The test information function is then defined in the natural way. For each latent trait we have

$$\mathcal{I}(\theta) = \sum_{i=1}^{n} \mathcal{I}_i(\theta). \tag{9.11}$$

In some applications, one may be interested in the standard error of estimation which is defined as the square root of the reciprocal of the test information function

$$SE(\theta) = \frac{1}{\sqrt{\mathcal{I}(\theta)}}. \tag{9.12}$$

Similar results hold for all the models presented earlier in this work. Bowing to brevity, we do not present all of these results. The interested reader may consult [10] for a more complete treatment.

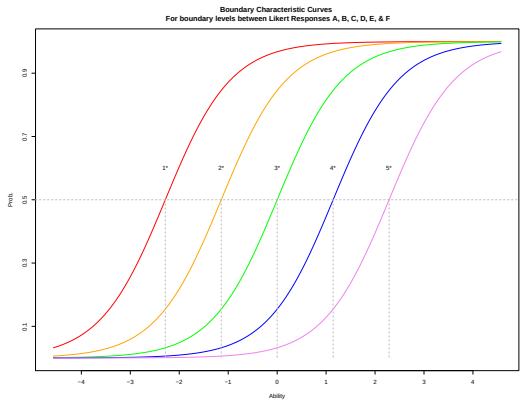


Figure 9.7: Example Boundary Characteristic Curves

Of particular interest for this work is the information coming from the graded response model of the previous section. In this case, we want to consider the boundary characteristic curves. In the case of the TTI Success Insights assessments, we typically have a Likert-style scale consisting of six possible categories as described in Section 9.3. Also noted in that section is that there exist five boundaries for the six categories. Each boundary curve represents the probability of a response at or above the lower category. As an example we plot a theoretical set of boundary characteristic curves (BCC) for one item for this type assessment.

Continuing to follow the development of the GRM model from Section 9.3, we define the item category response characteristic curves (IRCCC) as previously outlined. The following is a theoretical example based on the boundary curves presented in Figure 9.7.

It is noted that the information presented in Figures 9.7 and 9.8 are theoretically "ideal" in some sense. The curves are all equally spaced in both figures and the maximum values of probability of selecting a category are all equal for the four interior categories in Figure 9.8. It is unlikely to expect such high quality results in practice. We present them as examples to be used for comparison purposes.

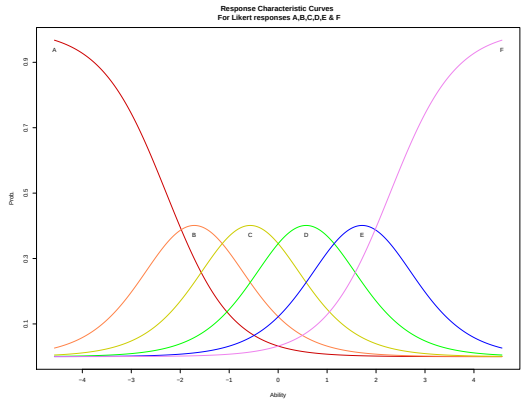


Figure 9.8: Example Item Response Category Characteristic Curves

We now turn our attention to constructing the item and test information functions for the polytomous response case, specifically for the GRM. Following [10], we note that the ultimate goal is to determine the precision with which a test or assessment measures ability and this has traditionally been measured via the test information function approach. It is also the case that the test information is a function of the individual item information functions, which are in turn functions of the category information functions. This may be seen as follows.

Note that we may consider the dichotomous case to be a simplification of the graded response model with only two cases. This is clearly true given that we model each boundary with the same functional approach as used in the dichotomous case, so the restriction is clear.

Define these two cases to be $P_{i1} = P_i$ and $P_{i2} = Q_i = 1 - P_i$ and denote by I_{ij} the amount of information associated with the j^{th} response category. We may then write the amount of information that is contributed by a given response category as the product of the probability of responding in that category and the score (information) one obtains given a response in that category. We may then sum this over all categories (in this case there are two) to obtain the item information,

$$\begin{aligned} I_i(\theta) &= I_{i1}(\theta)P_{i1}(\theta) + I_{i2}(\theta)P_{i2}(\theta) \\ &= I_{i1}(\theta)P_i(\theta) + I_{i2}(\theta)Q_i(\theta). \end{aligned}$$

From this point, the extension is straightforward, and we obtain for the case of n categories

$$I_i(\theta) = \sum_{j=1}^{n} I_{ij}(\theta)P_{ij}(\theta). \tag{9.13}$$

Relying on Samejima's development, see [154] and [155], define the information function of an item response category to be

$$I_{ij}(\theta) = -\frac{\partial^2 log(P_{ij}(\theta))}{\partial\theta^2} = -\frac{\partial}{\partial\theta}\left(\frac{P'_{ij}(\theta)}{P_{ij}(\theta)}\right), \tag{9.14}$$

where for brevity we have used the shorthand notation $P'_{ij}(\theta)$ to mean $\partial P(\theta)_{ij}/\partial\theta$.
The usual rules of differential calculus then give us that

TTI SUCCESS INSIGHTS®

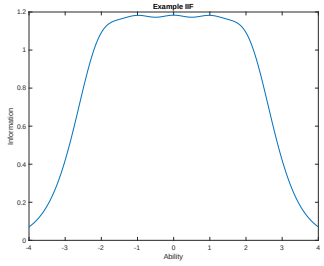


Figure 9.9: Example Item Information Function

$$I_{ij}(\theta) = \frac{\left(P'_{ij}(\theta)\right)^2 - P_{ij}(\theta)P''(\theta)}{(P_{ij})^2}. \tag{9.15}$$

We may then compute the share of the item information contributed by category j for item i to be

$$I_{ij}(\theta)P_{ij}(\theta) = \frac{\left(P'_{ij}(\theta)\right)^2}{(P_{ij})^2} - P''_{ij}(\theta). \tag{9.16}$$

Finally, the item information for item i becomes

$$I_i(\theta) = \sum_{j=1}^{n} \left(\frac{\left(P'_{ij}(\theta)\right)^2}{(P_{ij}(\theta))^2} - P''_{ij}(\theta) \right). \tag{9.17}$$

As a final note on item information, it is possible to formulate the entire development of the item information functions in terms of the boundary probabilities rather than the category probabilities. If we use the definitions of the P_{ij} in terms of the P^*_{ij} we are able to derive

$$\begin{aligned} I_i(\theta) &= \sum_{j=1}^{n} \frac{\left(P^{*\prime}_{ij-1}(\theta) - P^{*\prime}_{ij}(\theta)\right)^2}{P^*_{ij-1}(\theta) - P^*_{ij}(\theta)} \\ &- \sum_{j=1}^{n} \left(P''_{ij-1}(\theta) - P''_{ij}(\theta)\right). \end{aligned}$$

Of note, Samejima showed that the second summand in the previous equation vanishes identically, see [154]. In that case, the computation in terms of boundary probabilities reduces to

$$I_i(\theta) = \sum_{j=1}^{n} \frac{\left(P^{*\prime}_{ij-1}(\theta) - P^{*\prime}_{ij}(\theta)\right)^2}{P^*_{ij-1}(\theta) - P^*_{ij}(\theta)}. \tag{9.18}$$

TTI SUCCESS INSIGHTS®

Note that an interesting exercise is to show that (9.13) reduces to

$$I_i(\theta) = \frac{(P_i'(\theta))^2}{P_i(\theta)Q_i(\theta)} \tag{9.19}$$

if we assume the dichotomous case where $n = 2$ categories.

Now that we have the main pieces, we simply define the test information function to be the sum of the item information functions over all items of the assessment, e.g., from (9.17), we have

$$T = \sum_{i=1}^{N} I_i = \sum_{i=1}^{N} \sum_{j=1}^{n} \left(\frac{\left(P_{ij}'\right)^2}{(P_{ij})^2} - P_{ij}'' \right), \tag{9.20}$$

where we have suppressed the dependence on θ of the various functions. The final pieces are related to parameter estimation followed by ability estimation for the individual respondent.

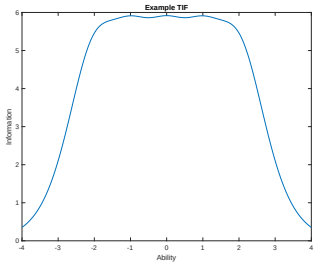


Figure 9.10: Example Test Information Function

Note that the authors of this manual have found little in the form of concrete statements as to what represents an acceptable level of information. In most literature reviewed for the writing of this manuscript the item information function is described as a measure of internal consistency reliability across the spectrum of ability (difficulty). It may then seem reasonable to apply the typical lower boundaries for reliability coefficients as first presented in Table 6.1 in Section 6. Recall that reliability coefficients with values in the range $0.70 \leq \alpha < 0.90$ are considered to be Respectable to Very Good. Values below 0.70 are Undesirable to Unacceptable and values above 0.90 may be considered to represent a highly homogeneous scale and one may want to consider shortening it. Note that we have used α to represent a generic reliability coefficient and not necessarily the traditional α coefficient mentioned in Section 6.

An example is given in Figure 9.11 showing an item information function along with a horizontal red line showing where the 0.70 value is achieved by the function. In this case, we see an item information function that appears to be strongly reliable throughout a large portion of the ability range, dropping off at the extremes of visible ability spectrum.

TTI SUCCESS INSIGHTS®

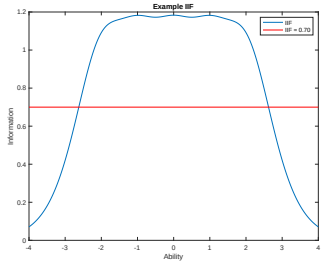


Figure 9.11: Example Item Information Function with Reliability Boundary

It seems a natural extension of the item information function to the test information function would be to determine the range of the ability spectrum for which the test information function exceeds $0.70n$ where n is the number of items that make up the assessment. In the example case at hand, we assumed there were 5 questions. Hence, we are interested in the portion of the ability range for which the test function exceeds $0.70 * 5 = 3.50$. An example of this is shown in Figure 9.12.

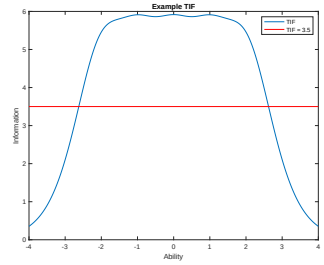


Figure 9.12: Example Test Information Function with Reliability Boundary

9.5 Parameter Estimation and Scoring Rubric

While the derivation of the parameter estimation and scoring that are an implicit part of the GRM modeling approach are interesting, constructive, and instructive, a full presentation is well beyond the scope of this work. Suffice it to say that the procedures are common, albeit lengthy and

TTI SUCCESS INSIGHTS®

somewhat challenging to follow for the non-mathematically adept individual. For now, we leave a more in-depth treatment of this topic to a future version of this work.

The importance of this topic, however, cannot be overstated. Computation of the parameters (α_i, and β_{ij} for each item i and each category j) allows for the GRM model to be operationalized in the form of analyzing the assessment items and the assessment itself, with appropriate assumptions regarding the population distribution of the ability parameter. In turn, once the model parameters are known, a scoring rubric may then be derived (using the same optimization technique, now with ability rather than the model parameters as the optimization variable(s)).

This optimization with respect to the ability variable may be done on an individual basis, thereby deriving a weighted scoring for the individual. It is not a difficult extension, then, to implement and utilize this scoring metric, assuming stable model parameters over large samples of the intended population. This is similar in nature to a factor analytic approach to scoring assessments in which the factor loadings across the items and their factorization may be utilized. In this way, one is lending more weight to the items deemed, by the underlying IRT or FA model, as more difficult and presumably such an item that should carry more weight in the scoring algorithm.

9.6 A Summary of the Graded Response Model Analysis

We have provided a general outline of the history and development of the item response theory approach to analyzing mental or psychometric based assessments. At this time, we use this approach as a diagnostic tool and note that the sheer amount of information present in this modeling approach really makes a full presentation of such result untenable. As an example, the TTI Success Insights EQ assessment consists of 57 questionnaire items, each of which would have an item response category characteristic curve and an item information function, combined with a test information function for each of the five constructs, and finally an overall test information function implies a total of 120 graphs that would need to be shown. A decision has been made to not include all of these graphics at this time. We do, however, present the test information functions by construct in Section 11.4.2. Historical results are presented in Section B.1.2. As is noted later in this work, as an organization, we have decided to use a weighted scoring model based on a factor analytic approach rather than an item response theory model.

10 Factor Analysis

The history and scope of modern factor analysis are well beyond the intent of the presentation of this topic in this manual. We choose to present a minimal amount of information that will hopefully allow the reader to gain a rudimentary understanding of the concepts. We attempt to lay out some basic historical and developmental facts as mentioned in many other works, e.g., [89], [95],[102], [103], and [125] provide excellent historical treatments. We also generously cite many researchers, both contemporary and historical, in an attempt to provide the interested reader with a comprehensive list of references for further study.

We begin the treatment by covering some preliminaries that are needed for the presentation of results in Section 11. This is followed by a presentation of some aspects of the historical treatment of factor analysis. We then discuss the geometric interpretation of correlation. After this, we present the basic factor model, the matrix form of the basic model, and the MinRes problem. We finish the section with a discussion of factor scores and indeterminacy and present a series of potential weighted scoring models that we use in our reporting system.

10.1 Some Preliminaries

Before the presentation of the meat of our factor analytic approach, we need to present some items that are used later in this work as part of our factor analysis results. As discussed elsewhere in this manual, the authors in both [45] and [74] note that while reliability is unquestionably important in assessing any assessment, validity is more important. Further, based on complications introduced by the attenuation paradox, see [45] and [109] among others, it may be worth sacrificing some reliability to acquire higher levels of validity during the assessment development process.

The full details of the factor analytic approach are beyond the scope of this technical manual. The interested reader may consult any of a multitude of resources on the subject. We provide a partial list of some of the authors' favorites on the topic, [19], [39], [66], [89], [102], [103], [104], [125], and [145]. We do present some basic definitions for use in the following analysis.

Definition 10.1. *If we denote by λ the vector containing the factor loadings for each item of a scale and further denote by $\mathbb{1}$ the vector of all ones of length n, where n is the number of items on the scale in question, then we may compute the composite reliability coefficient as follows:*

$$CR = \frac{(\lambda^T \mathbb{1})^2}{(\lambda^T \mathbb{1})^2 + 1 - \lambda^T \lambda}. \tag{10.1}$$

Further, we may compute the variance explained as

$$VE = \frac{\lambda^T \lambda}{n}. \tag{10.2}$$

The reader familiar with linear algebra likely associates the lower case Greek letter λ with one of its more common uses. In many settings, the symbol λ is used to denote the eigenvalues of a matrix. In other settings, λ may refer to the LaGrange multipliers associated with a constrained optimization problem. There is the possibility of confusing what the symbol is representing in a given situation. The context should make clear whether λ is referring to the factor pattern matrix entries (or factor loadings in a one factor model case as we deal with currently) or something else.

10.2 A Brief History of Factor Analysis

Most experts on the topic trace the origins of factor analysis to the turn of the 20^{th} century, specifically to the work of Charles Spearman. As noted in [95], while Karl Pearson is generally credited with setting forth the concept of principal axes, a central topic to the development of factor analysis, it is Spearman who is regarded as the "father of the subject". Thurstone generalized Spearman's tetrad-difference criterion to the rank of the correlation matrix as the basis for determining the number of common factors. According to [95], Thurstone equated Spearman's condition with the vanishing of determinants of matrices.

Hotelling, in the 1930s, capitalized on the work of Pearson and his ideas for principal axes and extended this concept into a full development of the method. However, given the time frame and the computational complexities involved, the work of Hotelling was premature in the sense of a clear need for high speed computing power to fully exploit the power of his work.

It appears that many of the results of this time foreshadowed the future use of high speed computing power to arrive at solutions to the problem of factor analysis. Many of the contributions to the research were mostly theoretical in nature, laying the groundwork for analyzing the factor solutions. As an example, various approaches were undertaken to view the problem of factor analysis from the viewpoint of a solution that would apply equally well to many areas of research (i.e., intelligence, personality, physical measurement, etc.) and any other variables of interest. An alternative viewpoint was to be concerned with "scientific meaningfulness rather than mathematical standard...", see [95] pp. 5.

According to [95] pp. 4,

> Thus, the chief aim is to attain scientific parsimony or economy of description.

On this topic of parsimony, [102] states what they call the *Postulate of Parsimony* as follows:

> This stipulates that, given two or more equally compatible models for the given data, the simpler model is believed to be true; in factor analysis, only the model involving the minimum number of common factors is considered appropriate.

The authors would like to point to the similarities between this principal of parsimony and the so-called Occam's, razor which states that of two explanations that account for all the facts, the simpler one is more likely to be the correct one. Thurstone proposed a *Simple Structure Theory* that is discussed in [102] as simply *Simple Structure* that they define as:

> A special term referring to a factor structure with certain simple properties; some of these properties include that a variable has factor loadings on as few common factors as possible, and that each common factor has significant loadings on some variables and no loadings on others.

A presentation of Thurstone's original ideas for simple structure are presented in [95] on pp. 97-98, where the author provides the following quotes from Thurstone's original work on the topic, see [171] for the original citation.

> If a reference frame can be found such that each test vector is contained in one or more of the ... coordinate hyperplanes, then the combined configuration is called a simple structure. (Thurstone 1947, pp. 328)

TTI SUCCESS INSIGHTS®

To provide more context for the above quote, the following is also presented on pp. 98 of [95].

> Just as we take it for granted that the individual differences in visual acuity are not involved in pitch discrimination, so we assume that in intellectual tasks some mental or cortical functions are not involved in every task. This is the principle of "simple structure" or "simple configuration" in the underlying order of a given set of attributes. (Thurstone, 1947, pp. 58)

The authors of this manual understand that the reader unfamiliar with factor analysis may not completely grasp the concepts being presented. Terms such as *simple structure* and *postulate of parsimony* are discussed further in subsequent sections when certain matrices integral to the factor analytic approach are presented. For now, we present a final topic related to historical development of factor analysis that may be viewed as controversial but is highly important nonetheless.

The concept of factor score is also discussed in a subsequent section. For now, we define a factor score simply to be a score derived from the results of a factor analysis based on a given set of underlying data. This factor score is, in some sense, a solution to a (constrained) optimization problem. The exact meaning of *optimal* is problem dependent and determined by one's choice of what one wishes to minimize. For now, we briefly discuss the concept of factor indeterminacy. In a later section, once some notation is available to us, we present the topic in more depth.

Underlying the factor analysis problem is, generally, a linear system of equations in some form. The problem of factor indeterminacy is related to the fact that the total number of factors (common and unique) exceeds the number of items. This in turn means that a certain matrix in the definition of the problem cannot be square, and hence cannot be inverted. The interpretation is that we have a system of equations in which we have more unknowns than we have equations. This implies that if a solution exists, then infinitely many solutions exist. The existence of an infinite number of solutions to a problem does not imply one cannot work with the system. It does, however, mean that one must establish further criteria for the existence of a specific solution. It also means that there are varying levels of consistency among the solutions. This may be interpreted in the sense that some solutions may have more determinacy than others, and the search is on for the best solution, whatever that may mean.

The next section lays out the general factor analysis problem and discusses some of the issues related to the problem. The section generally follows the first few chapters of [95], although other authors are consulted.

10.3 Geometric Interpretation of Correlation

Prior to any discussion on interpreting correlation, one should take the time to read the thoughts on current uses and interpretations of correlation in the social sciences we presented in Section 4.3. There appears to be a tendency for individuals in both the assessment industry and in the academy to over-generalize based on quite low levels of correlation. As discussed previously, one way to think of correlation is as the linear variation accounted for between the two variables. In this way, if we square the correlation coefficient we have an estimate of the variation accounted for between two variables. Having said that, note that it takes a correlation of at least $\rho = 0.7071068$ in order to claim that 50% of the variation between two variables is accounted for.

Before discussing the geometric interpretation of correlation, we must first introduce some notions from the branch of mathematics known as real analysis. First, let $a = (a_1, \ldots, a_n)$ denote an n-

dimensional vector. We usually write $a \in \mathbb{R}^n$. The (usual notion of) length of the vector a is given by the *Euclidean* norm which is defined to be

$$||a|| = \sqrt{\sum_{i=1}^{n} a_i^2}. \tag{10.3}$$

We may also talk about the scalar or dot product of two vectors. Let $a, b \in \mathbb{R}^n$ and define the scalar product to be $a \cdot b$ where

$$a \cdot b = \sum_{i=1}^{n} a_i b_i. \tag{10.4}$$

Note that when no confusion may arise, we write ab for $a \cdot b$. We need to establish a relationship between the vectors $a, b \in \mathbb{R}^n$ and the angle θ between them as is hinted at in Figure 10.1. If we imagine the triangle that is formed by the three vectors a, b, and $b - a$ we may write, using the so-called Law of Cosines from elementary trigonometry, a relationship that will allow us to prove the following.

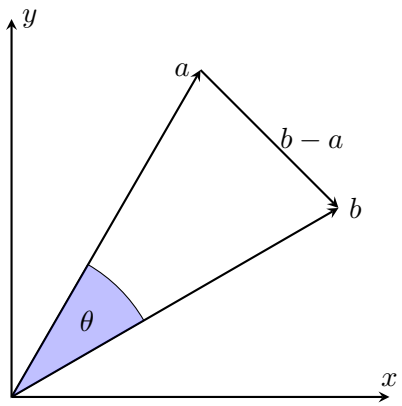


Figure 10.1: Vectors in $\mathbb{R}^n$

Proposition 10.2. *For any two vectors $a, b \in \mathbb{R}^n$, the angle θ between them is defined by the relationship*

$$\cos(\theta) = \frac{a \cdot b}{||a||||b||}. \tag{10.5}$$

Proof. As alluded to above, we shall employ a very unused relationship from elementary trigonometry thereby answering the age-old question of when will I ever use this in real life. So, here you go. The Law of Cosines states that the following relationship holds for the triangle presented in Figure 10.1,

$$||b - a||^2 = ||a||^2 + |b||^2 - 2||a||||b|| \cos(\theta). \tag{10.6}$$

Using both (10.3) and (10.4), applied to the left hand side of the equality in (10.6), we have

$$||b - a||^2 = (b - a) \cdot (b - a) = b \cdot b - 2b \cdot a + a \cdot a = ||b||^2 + ||a||^2 - 2a \cdot b, \tag{10.7}$$

TTI SUCCESS INSIGHTS®

where we have used the commutative property of the scalar product in $\mathbb{R}^n$. Substituting the result of the final equality of (10.7) into (10.6), we have

$$||b||^2 + ||a||^2 - 2a \cdot b = ||a||^2 + |b||^2 - 2||a||||b|| \cos(\theta) \tag{10.8}$$

which implies

$$\cos(\theta) = \frac{a \cdot b}{||a||||b||}, \tag{10.9}$$

the result claimed in (10.5) in the statement of the proposition. □

We now turn our attention to some notions of statistics, such as the familiar mean and standard deviation of random variables. Let X be a random variable and consider a random sample of X of size n. If we denote by p_i the probability of a random selection of item X_i, we may define the expectation of the random variable X to be

$$\mu(X) = E[X] = \sum_{i=1}^{n} X_i p_i. \tag{10.10}$$

Now, the reader may not be familiar with this particular form of the expectation of a random variable. In an elementary mathematics or statistics course, we usually assume in this setting that the random variables are, in some sense, uniformly distributed and we then have $p_i = \frac{1}{n}$ for all i. We then obtain the usual definition of mean or average as

$$\mu(X) = E[X] = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{10.11}$$

which is often denoted

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{10.12}$$

The other main statistic that most should be familiar with is a measure of the scatter or variation in a random sample X. We define the variance of X to be

$$\sigma^2(X) = E[(X - \mu)^2] = \sum_{i=1}^{n} (X_i - \mu)^2 p_i. \tag{10.13}$$

Note the use of the notation $\sigma^2(X)$. This is traditional and is a foreshadowing of the fact that most individuals are more familiar with the quantity $\sigma(X)$, the standard deviation of X, which is equal to the (positive) square root of the variance. If we adopt the same convention for p_i as in the definition of the mean, we have defined the biased estimator s^2 of $\sigma^2(X)$

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2. \tag{10.14}$$

Another possibility is to define $p_i = \frac{1}{n-1}$ and we have the so-called unbiased estimator

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu)^2. \tag{10.15}$$



Adopting the usual, but often incorrect, convention that "pretty much everything may be considered normally distributed", we may consider the standardized version of a random variable X defined as

$$x = \frac{X - \mu}{\sigma}, \tag{10.16}$$

and we may note that $\mu(x) = 0$ and $\sigma(x) = 1$. We state this in the following proposition.

Proposition 10.3. *Let X be a random variable with mean μ and variance σ^2. Let $Z = \frac{X-\mu}{\sigma}$. Then, we have*

$$\begin{aligned} \mu(Z) &= 0 \\ \textit{and} & \\ \sigma(Z) &= 1. \end{aligned} \tag{10.17}$$

Proof. We use the fact that the expectation operator is linear, i.e., that $E[\alpha X + \beta] = \alpha E[X] + E[\beta] = \alpha E[X] + \beta$, where α, β are scalars. So,

$$\begin{aligned} E[Z] &= E[\tfrac{X-\mu}{\sigma}] \\ &= \tfrac{1}{\sigma}E[X] - \tfrac{\mu}{\sigma} \\ &= \tfrac{\mu}{\sigma} - \tfrac{\mu}{\sigma} \\ &= 0. \end{aligned} \tag{10.18}$$

For the standard deviation of Z, it is convenient to first show that $\sigma^2(X)$ from (10.13) may be written in an alternative form.

$$\begin{aligned} E[(X-\mu)^2] &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned} \tag{10.19}$$

Now, using the final line of (10.19) and the results of (10.18) we have

$$\begin{aligned} \sigma^2(Z) &= E[Z^2] - E[Z]^2 \\ &= E[(\frac{X-\mu}{\sigma})^2] - 0^2 \end{aligned}$$

TTI SUCCESS INSIGHTS®

$$\begin{aligned} &= \tfrac{1}{\sigma^2}E[X^2 - 2X\mu + E[X]^2] \\ &= \tfrac{1}{\sigma^2}(E[X^2] - 2E[X]\mu + E[X]^2) \\ &= \tfrac{1}{\sigma^2}(\{E[X^2] - E[X]^2\} + \{E[X]^2 - 2E[X]\mu + E[X]^2\}) \\ &= \tfrac{1}{\sigma^2}(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2) \\ &= \tfrac{1}{\sigma^2} * \sigma^2 \\ &= 1. \end{aligned} \tag{10.20}$$

□

A brief note on an earlier comment, i.e., that everything is essentially considered normally distributed. This is actually not the case. In fact, most distributions are not normal. However, we, as a society, have a tendency to view things in "the average". Many incorrect assumptions are made about data and its usage, including incorrect interpretations of the Central Limit Theorem and the Law of Large Numbers. The Central Limit Theorem says, essentially, that if we have a random sample of size n taken from a population with mean μ and standard deviation σ, the limiting form of $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ is the standard normal distribution. This is a statement about the mean value of the random variable X, not about the distribution of X. The Law of Large Numbers simply says that the more times we perform the same experiment, the closer to the actual expected value we get. The idea here is that we do not have that all things are normally distributed. We do have that methods that work for normal distributions may work for non-normally distributed data under appropriate conditions.

Finally, we may look at how one data set or vector varies in relation to another. We have a natural measure between two separate data sets analogous to the variance of a random variable, and it is called covariation or covariance. We define

$$\sigma(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \frac{1}{m}\sum_{i=1}^{n}(X - \mu_X)(Y - \mu_Y). \tag{10.21}$$

We have used the notation $\mu(Z)$ to denote the mean of the random variable Z and m may be either n or $n-1$ based on whether one desires to use a biased or an unbiased estimate of $\sigma^2(Z)$. A natural extension of (10.21) is to normalize it via the following approach

$$\rho_{X,Y} = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}}. \tag{10.22}$$

If we standardize both X and Y in (10.22) using (10.16), we may write

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{10.23}$$

Now, from (10.3) and (10.4), we may write (10.23) as

$$\rho_{X,Y} = \frac{x \cdot y}{||x||||y||}. \tag{10.24}$$

Finally, from (10.5) we have

TTI SUCCESS INSIGHTS®

$$\rho_{X,Y} = \frac{x \cdot y}{||x||||y||} = \cos(\theta), \tag{10.25}$$

where θ is the angle between x and y in $\mathbb{R}^n$. To be explicit, the correlation coefficient between two data sets (vectors in $\mathbb{R}^n$) is defined by the cosine of the angle between those vectors. Note that $\cos(\theta)$ takes values in $[-1, 1]$, hence correlation between two vectors is bounded by the same.

We finish the discussion of correlation/covariance and its geometric interpretation by bringing into the mix the idea of statistical significance of a correlation computation and an application of covariation to the usual linear regression approach to modeling often used in social science settings. First, note that we often compute the statistical significance of a computation by using the so-called Student's test or Student's t-test. We do so here again. In this case, we compute the t-score as

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}}. \tag{10.26}$$

The null hypothesis in this case states that the correlation coefficient is equal to zero. If the hypothesis is false, then we may interpret this to mean that obtaining a correlation coefficient greater than ρ is highly unlikely. We thus compute

$$p = 2\int_0^t f(x)dx, \tag{10.27}$$

where t is the value computed in (10.26) and $f(x)$ is the probability density function of the t-distribution. For completeness,

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{10.28}$$

where ν represents the degrees of freedom and Γ is the gamma-function

$$\Gamma(n) = (n-1)! \tag{10.29}$$

for any positive integer n and

$$\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)dx \tag{10.30}$$

for any complex z with positive real part. Note that the t-distribution is part of most statistical packages and the information provided here is merely for completeness of exposition. However, this does allow one to compute a p-value that allows for an analysis of whether the correlation is statistically significant. Note that there is a difference between a correlation being statistically significant and a significant correlation between two variables. In one case, the statistically significant correlation simply means we should be able to trust the computed value as accurately describing any relationship between two variables. A significant correlation implies that a significant and potentially causal relationship actually exists between the two variables.

Finally, we look at the case of linear regression. Note that in most linear regression studies, one statistic is often reported as a level of the goodness of fit statistic. This statistic is often called the r-squared statistic, although it is really called the coefficient of determination. Suppose we have



a set of explanatory data X and a set of response data Y. Let $\hat{Y}$ denote the values of a linear fit model of the form $Y = aX + b$. We actually compute the r-squared statistic or coefficient of determination as the ratio of the variances between $\hat{Y}_i$ and Y_i as follows

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{10.31}$$

and we also consider the (positive) square root of (10.31)

$$R = \sqrt{R^2} = \sigma(\hat{Y}, Y). \tag{10.32}$$

10.4 The Basic Factor Model

We begin with a description of the basic factor model and some of its assumptions. In the social sciences, it is often assumed that one may be able to explain a large set of variables with a much smaller set of variables in a reasonable way. The basic idea is that we have observations of some variables, either latent or manifest. In either case, we have observations and we may then go onto hypothesize about a possibly more simple structure that may explain the observations in terms of hypothetical and unobservable latent variables called factors.

Perhaps noting an example from [103] that the reader may keep in mind for the discussions to come is in order. Suppose we have a series of observed variables of the form presented in Table 10.1. It is clear that these variables represent, in some fashion, political opinions that may be related to fiscal policy, social policy, liberalism versus conservatism, etc. The basic idea is that we do not need to attempt to define what they are related to. Rather, we make a basic assumption of the number of potential factors that may exist and then perform an exploratory factor analysis. Once we have established the factor analytic structure, it may then be possible to define what the factors are. The example in [103] posits the existence of two factors to explain this data, without naming these potential factors.

Table 10.1: Examples of Observed Variables from [103]

Variable	Definition
X_1	whether government should spend more money on schools
X_2	whether government should spend more money to reduce unemployment
X_3	whether government should control big business
X_4	whether government should expedite desegregation through housing
X_5	whether government sees to it that minorities get their respective quotas in jobs
X_6	whether government should expand the headstart program

The authors do state that, in their opinions, the first factor F_1 is responsible for the covariation in the first three variables and deals with economic issues. The second factor F_2 is responsible for the covariation in the last three variables and is related to civil rights issues.

The idea is to follow the lead of that discussed in Section 10.3 and create standardized variables Z_j rather than considering the set of X_j. The reason for this is pointed out in [95] pp. 12 where he



discusses the fact that the variables X_j are measured in arbitrary units with an arbitrary origin. Standardizing the variables then puts them all on the same scale with origin (mean) at zero with a unit variance, see Proposition 10.3. Once the variables are standardized, we may then form what [95] calls the *Full Classical Factor Model*,

$$z_{ji} = \sum_{k=1}^{m} a_{jk} F_{ki} + u_j Y_{ji}. \tag{10.33}$$

In (10.33) the term F_{ki} is the k^{th} common factor for the i^{th} individual, a_{jk} is the k^{th} factor loading for the j^{th} variable, Y_j is the unique factor for the j^{th} variable and u_j is its unique factor loading. The combination $a_{jk}F_{ki}$ for each of the m terms represents the contribution of the corresponding factor to the linear combination, while $u_j Y_j$ represents the residual error.

There are some traditional assumptions that need to be made clear. First, we assume that F_{ki} and Y_j are standardized variables, i.e., each has zero mean and unit variance. Further, the unique factors are uncorrelated with the common factors, usually denoted as

$$cov\left(F_{ki}, Y_{ji}\right) = 0, \quad \forall i. \tag{10.34}$$

Further, it is assumed that the unique factors are uncorrelated with each other,

$$cov\left(Y_{jk}, Y_{ji}\right) = 0, \quad \forall j. \tag{10.35}$$

If we apply (10.14) to 10.33, noting again that z_{ij} are standardized variables with zero mean and unit variance, we obtain

$$\begin{aligned} s_j^2 &= \sum_{i=1}^{N} z_{ij}^2/N \\ &= \sum_{k=1}^{m} a_{jk}^2 \left(\sum_{i=1}^{N} F_{ki}^2/N\right) + u_j^2 \sum_{i=1}^{N} \left(Y_{ji}^2/N\right) \\ &+ 2 \sum_{p<q=1}^{m} a_{jp}a_{jq} \left(\sum_{i=1}^{N} F_{pi}F_{qi}/N\right) + 2u_j \sum_{j=1}^{m} a_{jp} \left(\sum_{i=1}^{N} F_{pi}Y_{ji}/N\right). \end{aligned} \tag{10.36}$$

We interpret the terms in (10.36) as follows. The first line of the equation is the definition which is expanded in terms of the right hand side of (10.33). The second line contains two sums of sums. The parenthesized sums in each of the two outer sums are the definitions of the variance of the common and unique factors, respectively. Recall that both common and unique factors are standardize variables. The third line of (10.36) also contains two sums of sums. The internal parenthesized sums represent the covariance of the common factor with each other and the common factors with the unique factors. The latter is assumed to be zero. If we assume the former to also be zero, we have the so-called orthogonal factor case and we may write (10.36) as

$$s_j^2 = \sum_{i=1}^{m} a_{ji}^2 + u_j^2. \tag{10.37}$$

In (10.37), a_{ji} is the contribution of F_i to the variable z_j. We then define the total contribution of the factor F_i to all variables to be

$$V_i = \sum_{j=1}^{n} a_{ji}^2, \tag{10.38}$$

TTI SUCCESS INSIGHTS®

and further, define

$$V = \sum_{i=1}^{n} V_i \tag{10.39}$$

to be the total contribution of all common factors to the total variance. Then, V/n is the ratio defining the total variance explained. Clearly, one may consider V_i/n to be the contribution to the total variance explained of the i^{th} common factor. The term

$$h_j^2 = \sum_{i=1}^{m} a_{ji}^2 \quad j = 1, \ldots, n \tag{10.40}$$

is called the communality and the quantity

$$u_j^2 = 1 - h_j^2 \tag{10.41}$$

is called the uniqueness of the variable z_j. The variable u_j^2 indicates the extent to which the common factors fail to account for the total unit variance of the variable z_j. In some applications, we may wish to break up the unique factor as follows.

$$z_j = \sum_{i=1}^{m} a_{ji} F_i + b_j S_j + e_j E_j. \tag{10.42}$$

In (10.42), S_j and E_j are the specific and error factors respectively. We may then write uniqueness as $u_j^2 = b_j^2 + e_j^2$. Additionally, we may write the total variance as $s_j = h_j^2 + b_j^2 + e_j^2$.

The authors hope that at this point the reader has gained a feel for the approach taken to the basic factor analysis problem. It is not the intention of this section to completely present all aspects of factor analysis. Rather, we present a rough overview of the approach along with some tidbits of information that may be found to be useful. As an example of such useful information may be found in Table 10.2 where we have reproduced a similar table from [95]. The table synopsizes the information presented throughout this section of the manual. Prior to presenting this table, one more short definition and discussion is required. This presentation closely follows the discussion on pp. 19 in [95]. In this discussion, the term reliability is defined as the complement of error variance.

Table 10.2: Composition of Variance for Full Common Factor Model

Concept	Symbol	Definition	Alternate
Total Variance	1	$h_j^2 + b_j^2 + e_j^2$	$h_j^2 + u_j^2$
Reliability	r_{jJ}	$h_j^2 + b_j^2$	$1 - e_j^2$
Communality	h_j^2	$1 - u_j^2$	
Uniqueness	u_j^2	$b_j^2 + e_j^2$	$1 - h_j^2$
Specificity	b_j^2	$u_j^2 - e_j^2$	
Error Variance	e_j^2	$1 - r_{jJ}$	

In psychology, this systematic component of the variable (as distinguished from error components) is usually measured by the correlation between two separate administrations of the same test or parallel forms of the test.

TTI SUCCESS INSIGHTS®

In more contemporary terms, one measures reliability in the form of correlations based on test-retest or parallel forms reliability as defined, for example, by the APA in [74]. See also Section 6 in this manual for other discussions on the topic of reliability. If we define reliability to be the correlation between two different representations of the same variable, we then have z_j and z_J representing the two variables, and we define the correlation between them to be r_{jJ}, following the notation in [95].

10.5 Factor Analysis in Matrix Form

In this section, we use the more convenient notational form of linear algebra to condense the problem and simplify the presentation. In matrix form, (10.33) becomes

$$Z = Af + Uy, \tag{10.43}$$

where we choose to follow the notation of [95] again. In this setting, we have defined Z to be the complete set of N observations of each of the n variables. So, Z is an $n \times N$ matrix with entries $z^*_{ij} = z_{ij}/N$. The choice to divide by N is for a matter of notational convenience during matrix manipulation. Similarly, we define F to be the $m \times N$ matrix of common factors, Y is the $n \times N$ matrix of unique factors, and $(A|U)$ is the matrix A augmented with an $n \times n$ matrix containing the values u_j on the diagonal and zeros elsewhere.

If we denote by s_{ji} the correlation between the variable z_j and the factor F_i, we may define $s_{ji} = r_{z_j F_i}$ and further define $S = (s_{ji})$ to be the matrix of correlations between the standardized variables and the common factors. The matrix A from (10.43) is called the factor pattern matrix and the matrix S is called the factor structure matrix.

With the previous definitions at hand, we may derive a series of relationships between the factor pattern (A) and the factor structure (S). Begin with the following basic pattern relationship

$$Z = AF. \tag{10.44}$$

If we post-multiply (10.44) by F' (since F is not square, we must use the transpose) to obtain

$$ZF' = AFF' = S. \tag{10.45}$$

Note that the far left hand side of (10.45) is merely the definition of correlation between Z and F. It is also common to denote the matrix of common factor correlations by $\Phi = FF'$ and hence, $S = A\Phi$ is another way to express this relationship. This implies, assuming that Φ is non-singular, that $A = S\Phi^{-1}$. Recall that, by definition, the matrix of observed correlations is given by $R = ZZ'$. We may then represent the reproduced correlation matrix (i.e., the correlation matrix with communalities on the main diagonal) as

$$\tilde{R} = AFF'A' = A\Phi A' = S\Phi^{-1}S'. \tag{10.46}$$

10.6 A Condensed View of the MinRes Problem

This section provides a condensed discussion of the minimum residual, or minres, approach to the factor problem. The following quote from [95], pp. 174, is credited as being a large part of the foundation of solving the factor problem.

> The objective of a factor problem is to account for the tests, or their intercorrelations, in terms of a small number of derived variables, the smallest possible number that is consistent with acceptable residual errors.



This quote is from Thurstone in his 1947 work "Multiple Factor Analysis", see [171]. This approach is the alternative offered to, and not to be confused with, maximizing the variance extracted as in the principal factors approach. One may also consult any of a number of texts and articles on the topic, such as Chapter 8 in [95].

The basic idea behind principal factors or principal component analysis is a variable reduction based on maximizing the proportion of variance accounted for by each factor, successively. By this we mean that the first factor accounts for the maximum possible proportion of the variance that can be extracted by a single factor. The second factor then accounts for the maximum possible amount of the proportion of the variance that remains after the first factor has been extracted, and so on. The minres approach, rather than being concerned with accounting for variance of the original variables, attempts to reproduce the observed correlation matrix via the theoretically based correlation matrix with as little error as possible. This is accomplished via minimizing the off-diagonal sum of squares of the residuals between the observed and reproduced correlation matrices. Recall that the reproduced correlation matrix is an approximation of the observed correlation matrix with communalities on the diagonal.

This approach is described mathematically as follows. Recall that the factor problem may be written as $Z = Af + Uy$ where A is the factor pattern or factor loading matrix. We may attempt to find the best fit based on $\hat{R} = AA'$ where the matrix A is to be considered an unknown, $\hat{R}$ is the reproduced matrix, and R is the observed correlation matrix. We then employ the usual least squares approach to either fitting R by $\hat{R} + U^2$ or fitting $R - I$ by $\hat{R} - H^2$ where $H^2 = I - U^2 = diag(AA')$. In other words, H^2 may be recognized as the diagonal matrix with communalities on the diagonal.

Utilizing $\hat{R} + U^2$ leads us to the principal component approach. Using $\hat{R} - H^2$ and constructing an appropriate objective function leads to the minres solution. What is meant by an appropriate objective function is that we first desire, as mentioned earlier, to minimize the off-diagonal sum of squares of residuals between the observed and reproduced correlation matrices. Hence, we construct the sum of the off-diagonal residuals first utlizing

$$\min_{A} ||R - I - (AA' - diag(AA'))||. \tag{10.47}$$

Now, explicitly writing the objective function for the $\frac{n(n-1)}{2}$ off-diagonal residual correlations

$$f(A) = \sum_{k=j+1}^{n} \sum_{j=1}^{n-1} \left(r_{jk} - \sum_{i=1}^{m} a_{ji} a_{ki} \right)^2. \tag{10.48}$$

It is noted in [95] that the function $f(A)$ is dependent on the size of the correlation matrix. Hence, rather than minimize a quantity that necessarily grows with the size of the correlation matrix, we remove the dependency by scaling by the number of correlations to be computed. This new objective function is called the *root mean square* deviation and is given by

$$g(A) = \sqrt{\frac{2f(a)}{n(n-1)}}. \tag{10.49}$$

There is one further piece of the puzzle to discuss. We know from the earlier discussion, the communalities are directly computed during this process, see (10.47) and the fact that we are fitting $R - I$ by $\hat{R} - H^2$. By definition, communalities represent the amount of common variance shared by the factors and a given variable. This implies that it is not enough to minimize (10.48)

TTI SUCCESS INSIGHTS®

or (10.49). Rather, we must delve into the realm of constrained optimization. Surely it must be that a given communality h_j^2 must be bounded above by unity and hence

$$h_j^2 = \sum_{i=1}^{m} a_{ji}^2 \leq 1. \tag{10.50}$$

The authors do not intend to give a complete proof of the minimizer of (10.48) subject to the constraints (10.50), but we do present some of the details given that they are both instructive and constructive. The solution to this minimization problem is approached with the Gauss-Seidel method. Gauss-Seidel is an iterative technique from numerical linear algebra that essentially decomposes a matrix A into $A = L + U$ where L is lower triangular and U is upper triangular with zeros on the diagonal. The method then solves the problem $Ax = b$ by the iteration $Lx^{k+1} = b - Ux^k$.

First, it is noted on pp. 177 of [95] that the basic factor theorem given in equation (2.50) on pp. 32 of [95] implies that displacements in a single row of A introduce linear functions of the displacements in the reproduced correlation matrix. This further implies that the objective function is quadratic, at most. This is convenient in that it dramatically reduces the computational time given the computing power when this method was developed. Given modern computing power, this is really not very important. However, the approach to solving the problems of the time is instructive to the interested reader.

We first introduce a displacement in row j of A

$$A_j \mapsto A_j + \epsilon_j = (a_{j1} + \epsilon_1, \ldots, a_{jm} + \epsilon_m).$$

Next, we define a new variable $b_{ji} = a_{ji} + \epsilon_i$ and compute

$$\hat{r}_{jk} = \sum_{i=1}^{m} a_{ki} b_{ji}. \tag{10.51}$$

We now compute the j^{th} sum of squares of off-diagonal residuals as

$$\begin{aligned} f_j(A) &= \sum_{\substack{k=1 \\ k \neq j}}^{n} \left(r_{jk} - \sum_{i=1}^{m} a_{ki} b_{ji} \right)^2 \\ &= \sum_{\substack{k=1 \\ k \neq j}}^{n} \left(r_{jk}^* - \sum_{i=1}^{m} a_{ki} \epsilon_i \right)^2, \end{aligned} \tag{10.52}$$

where we have defined

$$r_{jk}^* = r_{jk} - \sum_{i=1}^{m} a_{ki} a_{ji} \quad \text{for} \quad k \neq j.$$

We now proceed to find the minimizer of (10.52) with respect to the displacement ϵ

$$\frac{\partial f_j}{\partial \epsilon_l} = 2 \sum_{\substack{k=1 \\ k \neq j}}^{n} \left(r_{jk} - \sum_{i=1}^{m} a_{ki} e_i \right) (-a_{kl}). \tag{10.53}$$

Finally, we set the right hand side of (10.53) equal to zero and solve for the terms containing ϵ_i. After some rearranging, which is justified as all sums are finite, we obtain



$$\sum_{\substack{k=1\\k\neq j}}^{n} r_{jk}a_{kl} = \sum_{i=1}^{m}\left(\sum_{\substack{k=1\\k\neq j}}^{n} a_{ki}a_{kl}\right)\epsilon_i. \tag{10.54}$$

The final solution is most readily demonstrated in matrix form as

$$\epsilon_j \bar{A}_j' \bar{A}_j = r_j^0 A \tag{10.55}$$

where we have defined $\epsilon_j = (\epsilon_1, \ldots, \epsilon_m)$, $\bar{A}_j'$ is the matrix A with the j^{th} row zeroed, and r_j^0 is the vector of residual correlations of variable j with all other variables and zero for the self residual. Now, under the assumption $\bar{A}_j'\bar{A}_j$ is invertible, we have

$$\epsilon_j = r_j^0 A \left(\bar{A}_j'\bar{A}_j\right)^{-1}. \tag{10.56}$$

We present a short comment on the invertibility assumption. It is not our intent to prove whether $\bar{A}_j'\bar{A}_j$ is invertible, but note that this issue was not addressed in the presentation in [95]. Given that $\bar{A}_j$ is defined as a matrix with all zeros in the j^{th} row, it is the opinion of these authors that a comment was in order in the original discussion.

A final note addresses the last issue, the so-called Heywood case and the reason for the constraints (10.50).

Theorem 10.4. *If the minimum of f_j is attained at a point outside the constraint set defined by (10.50), then a minimum of f_j subject to (10.50) will be attained on the boundary so that (10.50) may be replaced with*

$$\sum_{i=1}^{m} b_{ji}^2 = 1 \tag{10.57}$$

where we have replaced the a_{ji}^2 factor loadings with b_{ji}^2 as in (10.51) (since they are the minimization variables).

Theorem 10.4 is essentially the idea behind any constrained optimization on a compact set. For the reader unfamiliar with set theory, a simple way to understand compact sets is to keep in mind sets of real numbers that are both closed and bounded. The constraint (10.50) defines a compact set. The function (10.48) is clearly continuous. At the risk of over-simplification, this is essentially an application of the extreme value theorem (EVT), although the actual proof uses Lagrange multiplier theory rather than EVT. A discussion of the proof is provided in [95].

Given the discussion on the minres solution to the factor problem, it seems an appropriate time to re-introduce some geometric concepts of factor problems. Using the notation from throughout this section, we interpret the various pieces of Figure 10.2, in some sense, using the ideas of the minres solution to the factor problem. The vector z_j represents a response or an observation. This observation is projected onto a seemingly arbitrary plane defined by the factors, orthonormal in this case, F_1 and F_2. The distance between the vector z_j and its projection onto this plane, denoted $\hat{z}_j$, is the error ϵ_j. The coordinates of the vector $\hat{z}_j$ in the basis defined by F_1 and F_2 are given by the factor loadings and denoted a_{j1} and a_{j2}.

The variables and factors represented in Figure 10.2 may be thought of as representing the various entries of the matrices and objective function in the development of the solution to the minimum residual problem. In this figure, the minres solution is seeking to find, simultaneously, a solution



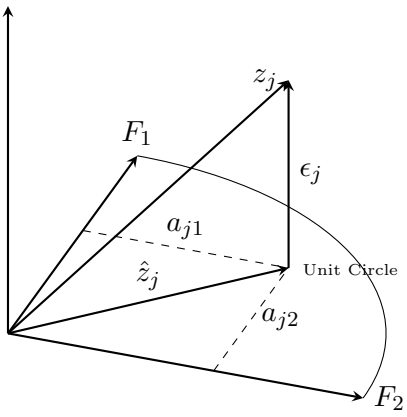


Figure 10.2: Geometric Interpretation of Factor Analysis Parameters

which minimizes the objective function given in (10.48) subject to the constraints (10.50) by finding that solution which produces the smallest residual errors ϵ_j. This is not a completely accurate comparison as the error ϵ_j in Figure 10.2 is the distance from the vector z_j and its projection onto the $F_1 - F_2$ plane. The actual minres approach minimizes the sum of the squares of the off-diagonal elements between the observed correlation matrix and the reproduced correlation matrix.

10.7 Factor Scores and Indeterminacy

In this section we discuss the problem of determinacy for the factor problem. Recall that in Section 10.5 we defined the matrix $(A|U)$ to be the matrix A augmented with an $n \times n$ matrix containing the values u_j on the diagonal and zeros elsewhere. For convenience, we now call this matrix $P = (A|U)$. Note that we many now write the system (10.43) in the form

$$Z = P(F|Y) = P\tilde{F} \tag{10.58}$$

where $(F|Y)$ is the common factor matrix augmented with the unique factor matrix. In theory, if P is invertible, a unique solution for the factor scores exists and is given simply by $\tilde{F} = P^{-1}Z$. However, the entire point of factor analysis, as discussed in the first part of Section 10.6 is to represent a test or its correlations with "few variables". This implies that the matrix P will rarely be square, the first requirement that P be invertible. This implies that the number of unknowns exceed the number of equations, further implying the possibility of an infinite number of solutions. This does not mean that factor scores cannot be computed. [92] provides the example of a criterion that must be satisfied for an orthogonal solution

$$\frac{1}{n}FF' = I_n, \tag{10.59}$$

which is the usual notation for an orthogonal matrix, and the matrix F is of size $n \times N$. Further, indeterminacy is not equal across all solutions, and hence, we may attempt to define solutions which are "better" than others based on criterion such as that just stated for orthogonal solutions. [92] also notes that relationships between factors and external criteria become more determinate as the factor scores become more determinate.

While the main resource for the factor score discussion is the paper by Grice, see [92], the notation is inconsistent, and there is a major typographic error in his equation (8). Well known author William Revelle notes the typographic error and has a brief discussion of the various factor score

TTI SUCCESS INSIGHTS®

approaches that may be read at https://personality-project.org/r/html/factor.scores.html. For now, we summarize Revelle's discussion using his notation.

Essentially, the problem boils down to finding an appropriate solution to

$$F = WZ, \tag{10.60}$$

or the score on factor j for person i may be written

$$F_{ji} = \sum_{k=1}^{N} w_{jk} z_{ik}. \tag{10.61}$$

The most simple form of a solution is to take $W = A'$ where A is the factor loading matrix from the P in (10.58). This is what Revelle calls his components method. An historical solution was proposed by Thurstone and is denoted $W = R^{-1}A$ where R is the observed correlation matrix and A is the factor loading matrix again. Harman proposes using $W = A(A'A)^{-1}$. The so-called Bartlett method uses $W = U^{-2}A(A'U^{-2}A)^{-1}$ where U is the diagonal matrix of uniquenesses. The Anderson method is only applicable to orthogonal factor solutions and is given by $W = U^{-2}A(A'U^{-2}RU^{-2}A)^{-1/2}$.

Finally, we have the ten Berge method. First, define the matrix $L = A\Phi^{1/2}$ where Φ is the matrix of factor correlations. Next, define the matrix $C = R^{-1/2}L(L'R^{-1}L)^{-1/2}$. Now, set $W = R^{-1/2}C\Phi^{1/2}$. There is clearly a range of differences in the various solutions for the matrix W in (10.60). These authors have no formulated opinion in favor of or against any of the methods. We present the material here as part of the larger overall discussion related to factor analytic procedures. We do want to stress an earlier remark based on information in [92]. Using factor scores should produce better criterion relationships than using other forms of scoring methodologies. The more determinate the factor scores are, the more likely one is to see more determinate relationships with external criteria. The main reason for this is the simple fact that not all assessment items are created equal, and the factor analytic approach allows for weighted scoring based on the relative strength (factor loads) and the levels of correlation that exist between the data and the factors, depending on the choice of method.

10.8 Weighted Scoring and Score Normalization

10.8.1 Rationale for Factor Analytic Weighted Scoring

Historically, assessment scores were computed using simple summation or arithmetic averaging of item responses within each construct. This approach assumes that every questionnaire item contributes equally to the measurement of the latent construct. While simple scoring methods have the advantages of transparency and ease of implementation, decades of psychometric research have demonstrated that questionnaire items rarely contribute equally to construct measurement.

Factor analytic methods provide empirical estimates of the relationship between individual questionnaire items and the underlying psychological construct they are intended to measure. Items exhibiting stronger relationships with the latent construct provide more information regarding an individual's standing on that construct, while items with weaker relationships contribute comparatively less information. Weighting item responses according to their empirically derived factor loadings therefore improves measurement precision by allowing more informative items to contribute proportionally more to the final construct score.

The weighted scoring methodology implemented within this assessment is based upon exploratory and confirmatory factor analytic procedures conducted during instrument redevelopment and val-



idation. The resulting factor score estimates more closely approximate the latent construct than simple arithmetic averages while maintaining the original conceptual framework and interpretation of the assessment.

Importantly, weighted scoring does not alter the theoretical definition of any construct. Rather, it improves the estimation of an individual's standing on that construct by accounting for the varying measurement properties of individual questionnaire items.

10.8.2 Score Normalization Through Power Transformation

Following implementation of weighted scoring, score distributions were evaluated across multiple normative samples. Consistent with man self-report personality and emotional intelligence instruments, positively worded questionnaire items produced distributions exhibiting moderate positive skew, resulting in construct means that were higher than theoretically desirable for interpretive purposes.

To improve score interpretability, a monotonic power transformation was applied to the weighted construct scores. The transformation preserves the relative ordering of respondents while moderately reducing score inflation within the upper portion of the distribution.

$$x^* = L + (U - L)\left(\frac{x - L}{U - L}\right)^{\lambda} \tag{10.62}$$

where L is the minimum possible score, U is the maximum possible score, and $\lambda > 0$ is a transformation parameter selected empirically.

Because the transformation is monotonic, respondent rankings remain unchanged. Individuals scoring higher than others prior to transformation continue to score higher following transformation. The transformation therefore affects score scaling rather than construct measurement. The primary effect of the transformation is a modest reduction in overall score means while preserving score ordering and construct representation.

Weighted scoring provides more accurate estimates of each emotional intelligence construct by reducing the influence of questionnaire items that contribute comparatively less measurement information. The resulting scores exhibit improved construct fidelity, improved discrimination among respondents, and enhanced stability across normative samples. Reduced score inflation also improves differentiation among individuals exhibiting moderate to high levels of emotional intelligence, facilitating more meaningful developmental conversations and reducing ceiling effects.

Because weighted scoring emphasizes the questionnaire items most strongly associated with each emotional intelligence construct, the resulting profile provides a more accurate reflection of the participant's underlying behavioral tendencies and emotional competencies. Score normalization likewise improves interpretability by reducing uniformly elevated scores while preserving respondent ordering and construct meaning.

Although the computational methodology has been enhanced, the underlying questionnaire content and construct definitions remain unchanged. The revised scoring methodology represents an improvement in measurement precision rather than a change in the theoretical framework of the assessment. Consequently, historical research conducted using previous versions of the assessment remains relevant to construct interpretation, while future administrations benefit from improved estimation of respondents' standing on each emotional intelligence dimension.

10.8.3 Benefits to Assessment Users

For assessment practitioners, coaches, consultants, and organizational decision makers, weighted scoring provides more accurate estimates of each emotional intelligence construct by reducing the influence of questionnaire items that contribute comparatively less measurement information. The resulting scores exhibit improved psychometric characteristics, including greater construct fidelity, improved discrimination among respondents, and enhanced stability across normative samples. These improvements increase confidence that observed score differences reflect meaningful differences in the underlying construct rather than artifacts of equal item weighting.sf

The score normalization process also produces distributions that more closely align with practitioner expectations regarding construct representation. Reduced score inflation improves differentiation among individuals exhibiting moderate to high levels of emotional intelligence, facilitating more meaningful developmental conversations and reducing ceiling effects frequently observed in self-report instruments.

10.8.4 Benefits to Assessment Participants

Assessment participants likewise benefit from these enhancements. Because weighted scoring emphasizes the questionnaire items most strongly associated with each emotional intelligence construct, the resulting profile provides a more accurate reflection of the participant's underlying behavioral tendencies and emotional competencies. Similarly, score normalization improves interpretability by reducing the tendency to self-report assessments to produce uniformly elevated scores. Participants therefore receive profiles that better differentiate relative strengths and opportunities for development across the various emotional intelligence dimensions. The improvements enhance the usefulness of the assessment as a developmental tool without altering the meaning or interpretation of the constructs being measured.

10.9 Exploratory Factor Analytic Summary

This section provides some historical motivation and some mathematical perspectives on exploratory factor analysis approaches to analyzing assessment data. The intent is to provide a synopsis of some of the literature while providing the foundation for why we are interested in such approaches. The actual results are presented in Section 11, where we present our most recent analytic results in Section 11.4.3 and our historical results in Section B.1.3.

11 Evidence of Validity Based on Internal Structure

This section is dedicated to presenting the results of our most recent analysis of our validity evidence related to internal structure. We present some motivation for why we present the areas we do study with the majority of the detailed information being presented in the relevant sections earlier in this work. We present corrected item-to-total correlation, item response theory, and exploratory factor analysis. We present the results in Section 11.4. We also present some of our historical results in Section B.1. We do note that, at this time we are not presenting results for the item response theory section. We perform the analysis and use it in our decision making process. However, the amount of information that is generated is quite large and prohibitive in the sense of being able to paint the full picture in this manuscript.

11.1 Corrected Item-to-Total Correlation

Corrected Item-total correlation (ITC) is more accurately called the corrected item to total score correlation. In other words, we are interested in determining to what extent individual item scores correlate with the total score on an aggregate basis.

The reason for looking at corrected item-total correlation is to obtain a better understanding of how well the individual items are measuring a given construct. Hence, item-total correlation may be thought of, in some sense, as a measure of evidence of the internal structure validity of a construct. If the item correlates well with the total score, it is measuring what is intended to some degree. Item-total correlation is not a replacement for other measures of validity of a construct. Rather, it is another piece of the story regarding the reliability and validity of an assessment.

Before discussing corrected item total correlation, we must first define item total correlation. If we let X denote the vector containing the total score of each respondent in a population on a scale and X_i denote the vector of individual scores of the respondents on the i^{th} item, then we have the vector of total scores defined by:

$$X = \sum_{i=1}^{n} X_i. \tag{11.1}$$

We may then define the item total correlation of the i^{th} item to be the usual correlation computation between X_i and X:

$$item\ total_i = cor(X_i, X). \tag{11.2}$$

As is seen in (11.1), the item total correlation between X_i and the total score X is dependent on the response to the i^{th} item. In order to better understand how the response to the i^{th} item impacts the total score, and hence how well the i^{th} item measures the underlying construct, it is necessary to remove this influence from the calculation in (11.2).

$$ITC_i = cor(X_i, X - X_i). \tag{11.3}$$

The calculation in (11.3) removes the influence of the i^{th} item and now provides a better estimate of the relationship between the response to the i^{th} item and the total score (with the i^{th} item removed). This, in turn, provides a better overall estimate of how well the i^{th} item measures the construct in question.



Table 11.1: Corrected Item-total Correlation Levels Suggested by APA

Range	Information Provided	Course of Action
< 0.1	Little to no information	Remove
$0.1 - 0.3$	Relatively small amount of information	Remove or revise
$0.3 - 0.5$	Moderate amount of information	Revise or retain as applicable
> 0.5	Large amount of information	Retain as desired

At the time of the writing of this manual, TTI Success Insights had not found a widely endorsed range of scores for item-total or average item-total correlations. However, some general guidelines state that scores between 0 and 0.19 indicate an item does not discriminate well. Values between 0.20 and 0.40 show a reasonable level of of discrimination. Finally, values above 0.40 show strong discrimination properties. Along a similar line of reasoning, [138], page 129, presents the information that is summarized in Table 11.1.

In addition, it is worth noting that item-total correlation scores that are too high, whatever that may mean, are potentially indicative of items/scales that are overly homogeneous and deserve further attention. For example, if a subset of items scores well above the average item total correlation score for a scale, one may want to highlight those items for further analysis in the factor analysis portion of the study as they may be redundant, more than one of the items providing little additional information.

In Section 11.4, results of the current corrected ITC are presented as part of the internal structure results. Item analysis is a quite encompassing topic given its relative simplicity of implementation. Several items are worth noting here. B.F. Green is quoted as stating that item correlations should be "a calm but insistent sea of small, highly similar correlations". His opinion is that the item correlation matrix should present as a series of very similar and non-volatile correlations. Put a different way, the correlation between any two items should not be far way from the mean item correlation value of all items of that scale, if they are to be adequately measuring the desired construct in a meaningful way.

This concept of staying near the mean value has a natural statistical measure called the standard deviation or the square root of the variance. It is then the desire for the item correlations, whether measured as item correlations, average inter-item correlations, or (corrected) item-total correlations, to have a solid mean with as small a standard deviation as is reasonable in a given situation.

11.2 Item Response Theory

We do not intend to rewrite the entirety of the information presented in Section 9. Rather, we will reiterate the reasons we consider item response theory approaches to analyzing psychometric instruments. Classical test theory is an approach that has been in use for nearly a century at this point. Item response theory is a modernized version that is intended to provide superior information about many of the concepts first studied under classical test theory.

For example, we are typically interested in the difficulty of particular items. Additionally, we are interested in how well the individual items discriminate. We should briefly discuss these topics in a bit more detail. Item difficulty literally determines how difficult an item is. In the sense of a correct versus incorrect testing approach, one computes the amount of the tested population that answers the item correct. In a psychometric instrument using a likert-style format, we are interested in determining how far from the median score the population mean falls. For example, on a 6-point likert scale with possible responses 1 through 6, the median response is 3.5. In a perfect world,



our item difficulty, or mean score, would be "close" to the median score for each of the items. One would accept some items that are above the mean (easier) and some below (more difficult), but would desire the majority to be close to the mean.

Item discrimination is a bit more difficult to describe. One is interested in how different individuals are with respect to the item difficulty. In other words, if an individual falls just slightly higher on the item difficulty, how much impact does this have on the level of the construct the individual possesses. This is a very esoteric concept when thinking in terms of the classical test theory approach. However, in the item response theory approach, item discrimination is closely linked to the steepness of the boundary functions that are described above in Section 9.

One of the beautiful things about item response theory is that we not only have natural connections to many classical test theory concepts of interest (the parameters of the item response theory functions directly give us difficulty and discrimination scores), the scores are dependent on each other. In classical test theory, we simply compute ratios of right to total or wrong to total, etc. In item response theory, we use an optimization algorithm to compute weighted versions of the statistics. In other words, the level of difficulty and discrimination of a given item is dependent on how the entire population responds to all items under consideration. This means that the difficulty and discrimination parameters take into account how difficult an item is or how well an item discriminates relative to all other items used to compute an individual's ability level on a given construct. This approach has much more utility and provides far superior information to that of classical test theory.

11.3 Exploratory Factor Analysis

Similar to the discussion in the previous section, we present more of the motivation for why we use factor analytic approaches rather than re-presenting the material from Section 10. Factor analytic approaches are similar to those of item response theory in that there is a natural relationship between the two. The main similarity is that the factor analytic approach generates a weighted system in a manner similar to that of item response theory. However, factor analysis has a feature that is more natural to the setting than that in the item response theory setting. Item response theory essentially assumes that everything is unidimensional, factor analysis does not.

We really like what item response theory brings to the table, but we want the flexibility of the factor analytic approach to generate multi-factor weighted scoring algorithms, which is not as natural to the item response theory approach. This is, at least for now, the experience these authors have had with item response theory. In the future, we hope to see research into multi-construct related item response theory functionality developed in more depth. The bulk of our approach to factor analysis is outlined in Section 10.

11.4 Internal Structure Current Results

This section contains the results of the current studies for corrected item-to-total correlation and exploratory factor analysis.

11.4.1 US Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the English-US version. The results are presented in Table 11.2. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table 11.2, we have highlighted potential problem areas in red. This item is identified for monitoring and possible replacement at some point in the future. There is no alarm to be sounded at this time. The mean item-to-total correlation scores are all in an acceptable range and the standard deviations show consistent scoring. The lower mean and larger standard deviation for Social Awareness is clearly due to the lone item scoring well below the mean. As of the writing of this manual, testing was underway for a replacement for this item. As soon as a suitable item is discovered, this item shall be replaced and all analyses updated.

Table 11.2: US Corrected Item-total Correlations 2026

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.41	0.60	0.57	0.50	0.41
Item 2	0.63	0.53	0.53	0.51	0.53
Item 3	0.69	0.44	0.53	0.46	0.61
Item 4	0.52	0.53	0.66	0.56	0.66
Item 5	0.73	0.62	0.58	0.49	0.45
Item 6	0.51	0.47	0.61	0.36	0.57
Item 7	0.73	0.49	0.50	0.18	0.63
Item 8	0.55	0.60	0.54	0.45	0.53
Item 9	0.75	0.56	0.57	0.45	0.60
Item 10	0.62	0.64	0.54	0.66	0.66
Item 11		0.56	0.58	0.65	0.54
Item 12		0.58	0.63	0.37	
Mean	0.61	0.55	0.57	0.47	0.56
StDev	0.11	0.06	0.05	0.13	0.08

11.4.2 US Item Response Theory

Section 11.2 above gives a brief discussion based on TTI Success Insights' application of the graded response models (GRM) we use in our item response theory (IRT) analyses. A more in depth discussion, along with myriad external references are provided in Section 9. In that section we present the main ideas behind generating a host of graphics to be used as part of our analyses. To be completely transparent, there are entirely too many potential plots to be generated than what could reasonably be presented in this work. There are boundary curves, item characteristic curves, item information functions, and test information functions, to name a few. We have chosen



to present the test information functions as a summary of the analyses performed. Any individual familiar with IRT in general, and GRM specifically, has the ability to interpret a great deal of information from a glance at the test information function plot of a GRM analyses.

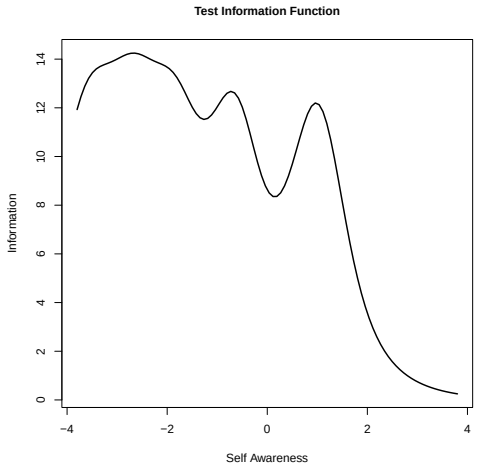


Figure 11.1: Self Awareness GRM TIF 2026

To be clear, the main immediate comparison is to the alpha coefficient from our internal consistency analyses. In this way, we are, on average, showing the range of the domain (of the construct) where the information meets or exceeds some minimum threshold. The interpretation is that we are measuring a continuous version of the point measure that is the α coefficient.

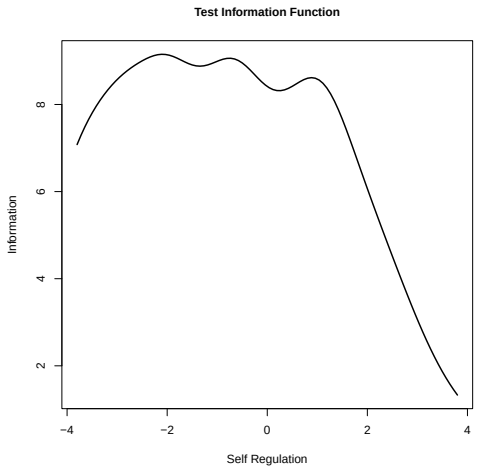


Figure 11.2: Self Regulation GRM TIF 2026

In some sense, we may interpret this as providing a measure of internal consistency at each point along the ability of our population across the continuum upon which it is defined. The larger the portion of the ability spectrum for which the test information function exceeds this level, the better the scale is performing. It should be noted that while our scores are reported on a fixed scale, say $[0, 100]$, technically speaking, the GRM approach assumes that the ability spectrum is, in fact, the entire real line from negative infinity to positive infinity. In practice, we usually see populations

TTI SUCCESS INSIGHTS®

fall between $[-4, 4]$ on the ability spectrum. We present the test information functions in Figures 11.1 to 11.5.

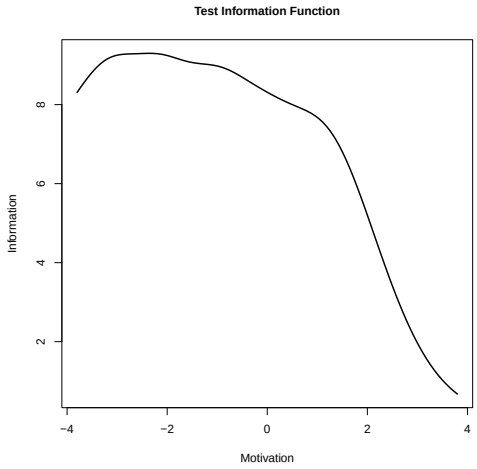


Figure 11.3: Motivation GRM TIF 2026

Self Awareness is presented first in Figure 11.1. This test information function shows a reasonable performance by this measure and is consistent with the findings in the internal consistency measurement presented in Section 6. Similarly, Self Regulation in Figure 11.2 is also showing reasonable levels of information. The two strongest performing scales based on test information are given by Motivation and Social Regulation, see Figures 11.3 and 11.5, respectively. The least well performing scale is given by Social Awareness in Figure 11.4. While it is the lease well performing, it still shows reasonable performance with a dip below the red line but remaining near it across a similar range as the others. We note here that similar to the item-total correlation results for Social Awareness, the lower level of performance shown in Figure 11.4 is largely due to the poor performance of a single item.

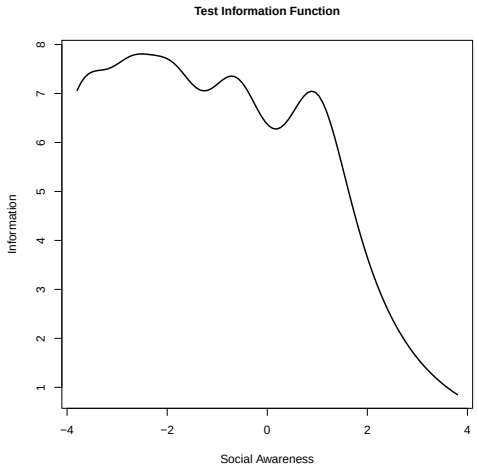


Figure 11.4: Social Awareness GRM TIF 2026

We now note that we have been discussing these models strictly from the standpoint of a comparison

TTI SUCCESS INSIGHTS®

to internal consistency, implying they may be better suited for presentation in Section 6, rather than in this section on internal structure. However, in a manner similar to the previous discussion, IRT in general, and GRM in this specific case, also provide a great deal of information about the item difficulty and discrimination. This is, again, to be interpreted as a continuous version of a point measure of the same concepts in classical test theory.

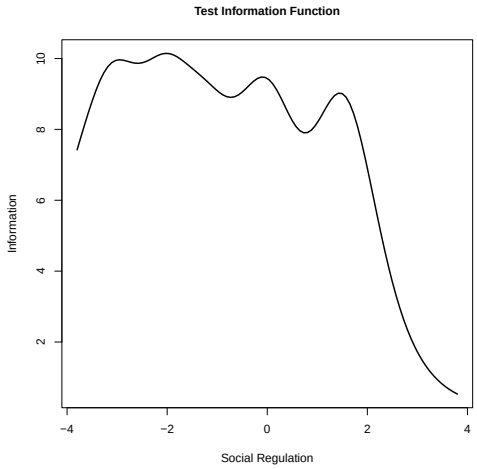


Figure 11.5: Social Regulation GRM TIF 2026

Test information, by definition, is the sum of the item information functions. Item information is relation closely to the derivatives of the item characteristic functions which are generated in the modeling process. Since item difficulty and item discrimination are inextricably linked to the item characteristic functions, and hence to the item and test information functions, the only way to have acceptable levels of information across the entire domain (i.e., across all difficulty levels, meaning across the ability spectrum), is for the item discrimination to function at an acceptable level across the difficulty spectrum. It then follows that acceptable levels of information are indicative of scales that are functioning at acceptable levels with respect to internal structure validity. In other words, we are establishing evidence of internal structure validity.

11.4.3 US Factor Analysis

This section presents the results of the factor analysis study for the general TTI population for the United States. The results are presented in Tables 11.3 through 11.7. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Some comments on how to interpret the information presented in Tables 11.3 through 11.7 are necessary. We are essentially looking for three main results. First, we wish to see a clean factor pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. The factor pattern matrices are the data presented in that aforementioned tables with any entries with magnitude less than 0.35 suppressed. We are essentially saying that any factor pattern matrix entry with a magnitude of less than 0.35 may be thought of as having little to no impact on a particular item/factor pairing.

Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it. We may read the tables either across the rows or down the columns. If we view it row by row, we are interpreting the impact any particular item has across

the factors. Ideally, any given item has a single factor on which it loads in an impactful way (i.e., above the 0.35 level). In some cases, we may find an item with loading on more than one factor and it may still be a perfectly acceptable model. If we view down the columns, we are identifying with items a particular factor is being loaded by.

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions. If we see any items with no loads above 0.35 we would identify that particular item for monitoring or replacement. If we have any factors with less that two items, the factor is not impacting the model in the intended manner and we would identify that factor for monitoring and action.

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good. The main takeaway for the US EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags.

There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation and improve any VE values to get them above the magic 0.50 value. This is the case for the Social Awareness construct where we see that VE has fallen to 0.47. This result is consistent with the historical information presented in Tables B.14 and B.15. In other words, the Social Awareness scale has proven to be difficult to improve based on the currently implemented approach.

We again note that this is largely based on a single underperforming item that has replacement in progress in the testing system employed by the TTI Success Insights continual improvement process. We further note that according to an internal study conducted, but not published, the average variance explained in the published literature for psychometric assessment scales was around 0.40. Given our scale is showing a variance explained value 15% higher than the average published value, we feel solid in continuing to use the questionnaire, as is, while we continue our pursuit of more appropriate items to replace the underperforming item on this scale.

Table 11.3: US Factor Pattern Matrix
Self Awareness 2026

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.43	
Item 2	0.66		
Item 3	0.64		
Item 4			0.57
Item 5	0.63		0.38
Item 6		0.59	
Item 7	0.66		
Item 8			0.61
Item 9	0.62	0.37	0.36
Item 10		0.74	
VE	0.53		

Table 11.4: US Factor Pattern Matrix
Self Regulation 2026

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.36		0.41	
Item 2				0.76
Item 3			0.60	
Item 4			0.64	
Item 5		0.78		
Item 6			0.46	
Item 7	0.72			
Item 8				0.69
Item 9	0.81			
Item 10		0.81		
Item 11				
Item 12	0.58		0.35	
VE	0.55			

Table 11.5: US Factor Pattern Matrix
Motivation 2026

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.80		
Item 2		0.76	
Item 3		0.69	
Item 4	0.83		
Item 5	0.69		
Item 6	0.63		
Item 7			0.67
Item 8			0.49
Item 9			0.70
Item 10		0.70	
Item 11		0.39	0.51
Item 12	0.49		0.44
VE	0.55		

Table 11.6: US Factor Pattern Matrix
Social Awareness 2026

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.58	
Item 2	0.79			
Item 3	0.73			
Item 4		0.54		
Item 5		0.43		0.35
Item 6		0.63		
Item 7	0.39			
Item 8			0.83	
Item 9	0.36			
Item 10				0.53
Item 11		0.50		0.52
Item 12		0.47		0.46
VE	0.47			

Table 11.7: US Factor Pattern Matrix
Social Regulation 2026 Current

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.84
Item 2		0.63		
Item 3			0.80	
Item 4			0.38	
Item 5				0.83
Item 6	0.79			
Item 7	0.79			
Item 8		0.84		
Item 9	0.52			
Item 10	0.37		0.79	
Item 11		0.58		
VE	0.65			

TTI SUCCESS INSIGHTS®

11.5 Summary

This section presents the results of the TTI Success Insights validity evidence based on internal structure. Our analyses include corrected item-to-total correlation (CITC), item response theory (IRT), and exploratory factor analysis (EFA). While each approach generally speaking provided the same or, at least, very similar information, each is presented for particular reasons.

The mathematically challenged reader may not be completely comfortable reading the information that is provided in the test information functions from IRT. It is also somewhat esoteric to view the results of the EFA and have a complete understanding of what they all mean. For these reasons, we first present the correlation results above. As always, we emphasize the fact that no single mathematical or statistical analyses provides a complete argument establishing evidence of internal structure validity. We present the historical analytical results in Appendix B below.

12 Evidence of Validity Based on Relationships to External Variables

The American Psychological Association (APA) identifies five key areas of any psychometric assessment requiring evidence of validity, see [163]. One of these areas is *Relationship to Other Variables*, which encompasses several concepts such as criterion, concurrent, and predictive validity. While [75] presents many of the theoretical concepts of evidence of reliability and validity of assessments, [4] is an effort by the Joint Committee on Standards for Educational and Psychological Testing to describe, in more detail, guidelines related to these ideas.

As an example, Standard 1.4 of [4] states

> If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary.

There are at least two major takeaways from the previous statement. The first is more implicit in that any use of assessment data in a way that has not been validated requires validation. Second, and more explicit, is the responsibility for providing evidence of such validation lies primarily with the test user.

One of the purposes of this paper is to begin to establish a foundation of validity evidence for the TTI Success Insights EQ assessment by combining demographic information, specifically job description, with a classification algorithm. In this case, the authors choose to use a logistic regression approach to assigning membership in a group. One may consult Section 4.3 above for a more detailed discussion presenting some of the reasons for our choice.

In a previous study, [73] shows that the TTI Success Insights TriMetrix DNA Legacy variables may be used in conjunction with a logistic regression classification approach to successfully differentiate serial entrepreneurs from a group randomly selected from the TTI Success Insights database. This study is an example demonstrating the relationship between TTI Success Insights assessment derived variables (Behaviors, Motivators, and DNA) with other variables (membership in a specific population within a larger population).

In April, 2018 TTI Success Insights began offering the opportunity for our network of Value Added Associates (VAA) to add demographics collection to their assessment links. Since that time, a large amount of demographic information has been collected and analyzed in various ways. The current study under consideration asks whether the use of a classification technique in a manner similar to the previously discussed paper, see [73], may be applied to subsets of the demographic data. In particular, this study considers several subsets defined through the use of the "Job Description" demographic category in the TTI Success Insights demographics database.

12.1 A Brief Review of Logistic Regression

The current study breaks each data set into two subsets, the Target group with classification equal to 1, and the Control group with classification equal to 0. In other words, our classification variable is binary. Note that one may consider more than two classifications using a multiple logistic regression approach.

Generally speaking, there are a multitude of excellent references on logistic regression. This paper follows the work in [97], but also relies on the work in [2]. Suppose we have a single response variable

y taking values in $\{0, 1\}$ and a single, continuous explanatory variable x. The corresponding logistic regression model is of the form

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{12.1}$$

where the notation exp denotes the usual exponential function with base e. The function $\pi : D \mapsto [0, 1]$ where D is an appropriate domain dependent on the explanatory variable x and $[0, 1]$ is the usual unit interval in $\mathbb{R}$.

According to [97], there are two main reasons for choosing the logistic distribution in (12.1). First, π is an extremely flexible and easily used function, and second, π lends itself to meaningful (clinical) interpretation. To see the utility of the function π, note the following transformation, called the logit transformation,

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right). \tag{12.2}$$

Note that with a little algebra, $g(x) = \beta_0 + \beta_1 x$. This is useful in that the logit transformation of the logistic regression equation results in a linear expression with many of the desirable properties of the usual linear regression model.

One important difference between linear and logistic regression is that the error, which expresses an observation's deviation from the conditional mean, is no longer assumed to be normally distributed. Again following [97], we may express the value of the outcome variable given x as $y = \pi(x) + \epsilon$.

In this formulation, ϵ may take on one of two possible values. If $y = 1$, then $\epsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\epsilon = -\pi(x)$ with probability $1 - \pi(x)$. In summary, ϵ follows a binomial distribution with probability given by the conditional mean $\pi(x)$.

The importance of the preceding discussion is that we can now readily construct the likelihood function of the above mentioned binomial distribution. For values of $y = 1$ given x the contribution to the likelihood function is $\pi(x)$ and the contribution for values of $y = 0$ given x the contribution is $1 - \pi(x)$. Thus, for any observation x_i, the contribution to the likelihood function is given by

$$\pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1 - y_i}. \tag{12.3}$$

Note that (12.3) reduces to $\pi(x_i)$ or $1 - \pi(x_i)$ depending on the value of y_i given the choice of x_i. One assumption in logistic regression is that the observations are independent and, hence, the likelihood function is given by the product of the individual terms given in (12.3):

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1 - y_i} \tag{12.4}$$

There is one more step involved to obtain the desired result. In all parametric regression approaches, there is an underlying optimization. This usually entails some form of differentiation. In the case at hand, (12.4) now requires differentiation with respect to the parameters $\boldsymbol{\beta}$ and a solution of the resulting equations. However, differentiation of products of functions is quite difficult compared to differentiation of sums of functions. This leads to a heavy computational cost. Hence, it is advantageous to construct the log-likelihood function by taking the logarithm of (12.4) and using the appropriate properties of the logarithmic functions, namely that $\ln(f \cdot g) = \ln(f) + \ln(g)$ and $\ln(f^g) = g \ln(f)$. Hence,

TTI SUCCESS INSIGHTS®

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \{y_i \ln(\pi(x_i)) + (1 - y_i)(\ln(1 - \pi(x_i)))\}. \tag{12.5}$$

defines the log-likelihood function.

The problem at hand is now to optimize (12.5) with respect to the parameters $\boldsymbol{\beta}$. It should be noted that while the parameters $\boldsymbol{\beta}$ are not explicitly present in (12.5), one may substitute the definition of $\pi(x)$ from (12.1) into (12.5) to see that (12.5) is, in fact, a function of the parameters $\boldsymbol{\beta}$.

An extension of logistic regression that may be useful in classification problems is that of multinomial logistic regression. As a quick example, suppose that the response variable now may take on any of three possible values, $\{0, 1, 2\}$. In this case, one may define the conditional probabilities of each outcome category as follows:

$$P(y = 0|x) = \frac{1}{1 + \exp(g_1(x)) + \exp(g_2(x))}, \tag{12.6}$$

$$P(y = 1|x) = \frac{\exp(g_1(x))}{1 + \exp(g_1(x)) + \exp(g_2(x))}, \tag{12.7}$$

and

$$P(y = 2|x) = \frac{\exp(g_2(x))}{1 + \exp(g_1(x)) + \exp(g_2(x))}, \tag{12.8}$$

where

$$g_i(x) = \beta_{i0} + \beta_{i1}x_1 + \ldots \beta_{in}x_n. \tag{12.9}$$

In (12.9) the index i runs from 1 to the number of categories present (2 in this example), and n represents the number of independent variables present.

There is a similar derivation of the log likelihood function to that in (12.5) and a maximum likelihood estimation process is used to find the coefficients (β_{ij}).

The utility of the multinomial logistic regression technique is for a case similar to predicting the likelihood of a student with a given set of characteristics to pass a given course with a particular grade level. This process could also be useful in constructing a predictive model that would rank a group of sales employees into two categories, one category representing high performers and the other category representing low to average performers. The third category may be a random sample of the general population for differentiation purposes.

12.2 Confusion Matrices, ROC, and AUC

This paper relies on the concepts of the receiver operating characteristic (ROC) curve and various derived statistics that are based on the information contained in the confusion matrix. For a more detailed version of the content of this section, one may consult [69]. For some critiques of the approach with additional options offered, one may be interested in [140].

The confusion matrix is an organizational tool that contains information on the correct and incorrect classification counts of the model of interest. The goal here is to predict membership in a group.



		Prediction outcome	
		p	**n**
actual value	**Y**	True Positive	False Positive
	N	False Negative	True Negative
	total	P'	N'

Figure 12.1: Confusion Matrix

We may classify a prediction based on whether it correctly identifies membership. In this way, we wish to know the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

A TP is assigned to a prediction if the prediction assigns membership to the group of interest, and the individual belongs to the group of interest. FP is assigned to a prediction that assigns membership to the group of interest, and the individual does not belong to the group of interest. Similarly, TN is assigned to a prediction if the prediction assigns membership to the alternate to the group of interest, and the individual does not belong to the group of interest. Finally, FN is assigned to a prediction if the prediction assigns membership to the alternate to the group of interest, and the individual belongs to the group of interest.

We organize this information in the confusion matrix as follows. Let **p** and **n** denote the predicted outcomes and let **Y** and **N** denote actual membership. Further, let P' and N' denote the total positive and negative predictions, respectively. In other words, P' is the sum of TP and FP while N' is the sum of FN and TN. Figure 12.1 displays the resulting information contained in a confusion matrix. Note that $Y = P'$ and $N = N'$, by definition. In other words, the sum of TP and FN is the actual number of members in the group of interest and similarly for the sum of FP and TN.

What has not been specified is how one determines whether a prediction correctly identifies membership. To address this, one must first define at what level one considers a prediction to be correctly classifying. For example, suppose we have 100 individuals and 25 of them belong to a group of interest. A reasonable cutoff would be to state that if the model assigns a probability of belonging to the group of interest greater than or equal to 0.25, we would consider that to be a positive. If the individual with a probability of at least 0.25 actually belongs to the group of interest, we increment the TP counter by one, otherwise we increment the FP counter by one. Similarly for a probability less than 0.25 and the negative counters.

It is also reasonable to consider the cutoff of 0.50, especially if one knows that the sample population does not accurately depict reality. The fact is that it is highly unlikely for the modeler to know what the true population breakdown is. Therefore, we consider many different cutoff points which leads to the concept of the ROC curve. We first need to define some of the essential statistics, some of which are directly used in this paper, and others are for information only.

First we define the true positive rate and the false positive rate. The true positive rate is the ratio

of correctly identified positives to the total number of actual positives. In terms of the confusion matrix we have

$$TPR = \frac{TP}{P'}. \tag{12.10}$$

Similarly, the false positive rate is the total number of false positives divided by the total number of negatives.

$$FPR = \frac{FP}{N'}. \tag{12.11}$$

Some quantities of interest may be accuracy, the total correct classifications divided by the sample size:

$$Acc = \frac{TP + TN}{P' + N'}, \tag{12.12}$$

precision, the true positives divided by the total number of positives, true and false:

$$Pre = \frac{TP}{TP + FP}, \tag{12.13}$$

and a measure of overall fit called the F-measure

$$F = \frac{2}{\frac{1}{Pre} + \frac{1}{Acc}}. \tag{12.14}$$

Going back to the example mentioned before, we have a group of 100 total, and 25 we are interested in identifying using some method. Suppose we have the following confusion matrix:

Table 12.1: Example Confusion Matrix

17	12
8	63

We may compute the above mentioned statistics.

$$TPR = \frac{17}{25} = 0.68, \tag{12.15}$$

$$FPR = \frac{12}{75} = 0.16, \tag{12.16}$$

$$Acc = \frac{17 + 63}{100} = 0.80, \tag{12.17}$$

$$Pre = \frac{17}{17 + 12} = 0.59, \tag{12.18}$$

and

$$F = \frac{2}{\frac{1}{.59} + \frac{1}{.8}} = .68. \tag{12.19}$$

TTI SUCCESS INSIGHTS®

We note that the above example is for a single paring of TPR with FPR. Each of these rates has been computed for a single classification point, the examples given were 0.25 and 0.50 for the fictitious data in Table 12.1. The previous quantities are not based on a real classification model.

To move to the concept of the ROC curve, we consider not just a single point, but all possible values we could use as a judgment of how well the model performs at each possible value in the interval $[0, 1]$. Since there are infinitely many possibilities, we take a reasonably sized discrete sample and compute a confusion matrix for each value we choose.

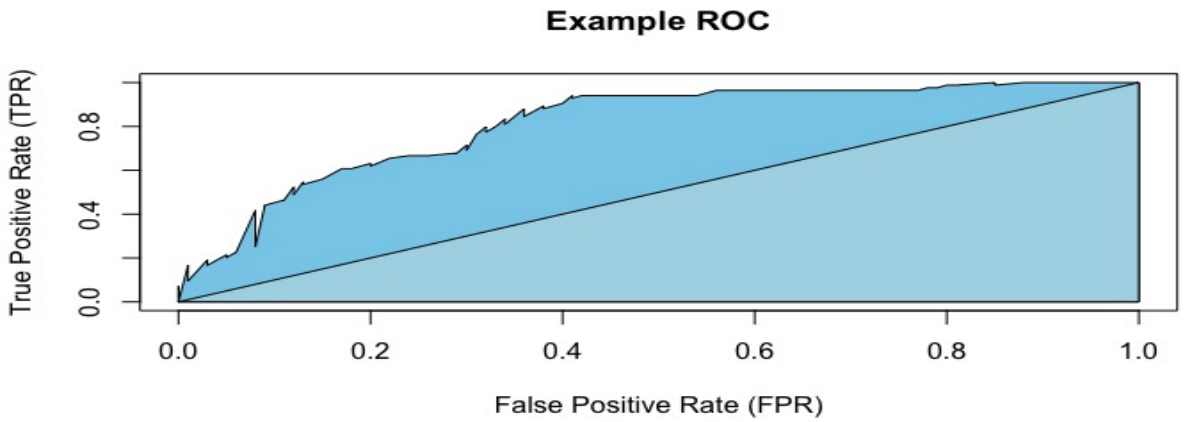


Figure 12.2: Example ROC Curve

For each confusion matrix, we may compute any or all of the aforementioned quantities. However, we are particularly interested in the relationship between the true positive rate and the false positive rate. We consider how the functional relationship behaves, TPR as a function of FPR by considering these as ordered pairs of data (FPR_i, TPR_i), $i = 1, \ldots, n$, in the unit square in $\mathbb{R}^2$.

Visualization of this data is done by plotting the ordered pairs of data with the FPR_i values along the x-axis and the TPR_i values on the y-axis. This plot is known as the ROC curve. Figure 12.2 shows an example plot. The line separating the two shades of blue is the line $y = x$ and represents the ROC curve of a random selection process. In other words, if we randomly classified individuals, the line $y = x$ represents the relationship between the FPR and TPR values.

Values of an ROC curve that lie above the line $y = x$ denote a classification algorithm that is superior to a random selection process with the reverse holding for points below the line. The example in Figure 12.2 shows a solidly performing algorithm. It is desirable to have a quantification of how much better the algorithm performs than the random approach. This value is given to us by the area under the curve (AUC), meaning the area under the classification ROC curve.

The line $y = x$ forms a right triangle with the x-axis and the vertical line $x = 1$. The area under this curve on the interval $[0, 1]$ is thus 0.50. If we can show the area under the ROC curve of interest is greater than or equal to 0.50, we have shown that the classification algorithm in question performs better than the random selection approach.

The solution to this problem is to integrate the function defining the ROC curve. In other words,

$$AUC = \int_0^1 f(x)dx \tag{12.20}$$

where $f(x)$ is a functional representation of the FPR-TPR relationship. In particular, we have that

$$AUC = \int_0^1 f(x)dx = \frac{1}{1-0}\int_0^1 f(x)dx \tag{12.21}$$

which shows that AUC is the average value of the function $f(x)$ on the interval $[0, 1]$.

We do not have an explicit form for $f(x)$ generally speaking, but we do have techniques to integrate discrete functions that we may apply. For one such technique, the interested reader may consult Algorithm 1 in [69].

12.3 A Brief Introduction to Principal Component Analysis

This section provides a brief review of principal component analysis and is based mostly on the second chapter of [64].

Generally speaking, both the Style Insights and Motivation Insights portions of the Talent Insights assessment are forced-choice assessment questionnaires. The following fact is true of all forced-choice assessments. If we set r to be the number of items in a forced-choice block on an assessment, then the off-diagonal elements of the correlation matrix of the scales of the forced-choice assessment are given by

$$cor(X, Y) = -\frac{1}{r} \tag{12.22}$$

where X and Y represent any two scales measured by the assessment.

In a given modeling exercise, it is desirable to use explanatory variables that are correlated with the response variable, but not with other explanatory variables, in order to avoid overfitting. Given that the assessment scales are naturally correlated by the nature of the assessment, it may be possible to retain the information provided by the explanatory variables but do so in a way that removes as much correlation as possible.

One way of approaching this problem is to consider principal component analysis (PCA). PCA is a variable reduction process that uses linear combinations of the original variables to create a, presumably smaller set of variables that accounts for as much of the variation in the original variables as possible.

Mathematically, if we let X_i, $i = 1, \ldots, n$ denote the set of original variables we define

$$PC_1 = \sum_{i=1}^{n} w_{(1)i} X_i \tag{12.23}$$

where the weights $w_{(1)i}$ are chosen to maximize the ratio of the variance of PC_1 to the total variation subject to the following constraint

$$\sum_{i=1}^{n} w_{(1)i} = 1. \tag{12.24}$$



We then compute the second new vector PC_2 orthogonal to PC_1 which accounts for the maximum amount of the remaining total variation not accounted for by PC_1. This process continues until we have computed n new vectors that form an orthogonal basis for the space spanned by the original set of n explanatory variables. At this point, the art in the phrase "art and science" comes into play.

We choose the first m new variables $PC_1, \ldots, PC_m$ such that as much variation as possible is accounted for in the following sense. When adding the next variable to the set to be considered no longer adds a significant amount of variation accounted for, that next variable is not adding to the overall information to be considered. Exactly what the cutoff should be is a matter of some trial and error to see what works best in a given situation.

One positive takeaway from the PCA approach is that we now have a set of orthonormal variables and the probability of overfitting the model is much smaller. A negative of this approach is that the interpretation of the new variables is not necessarily straightforward.

12.4 Logistic Regression Study Results

Since 2018 the TTI Success Insights Research and Product Development team has been collecting certain demographic data from our questionnaire respondents on a voluntary basis. One of the categories we collect is the respondent's job, when applicable. We are using the O*Net job classification system. The respondent is offered the opportunity to choose both their industry and job during this process.

A decision is made to consider only those occupational titles for which TTI Success Insights has 100 or more respondents in our database. In the event that an occupational title has more than 500 respondents, a decision is made to incorporate a holdout sample testing approach. The holdout sample approach employs five randomly created subsets of the data. The modeling exercise was then applied to the data in such a manner that five separate models were created. Each of the models uses a different subset of four of the five data samples as training data. The results of the training model are then applied to the holdout sample and the results recorded.

The intent of this approach is to not only identify the statistically significant variable or variables that allow for identification of the target population, but to do so on multiple, independent data sets. In this way, it is possible to establish that the relationship is more than just a spurious one. If the same variable or variables consistently come through the modeling process as statistically significant, the relationship may be considered more robust than if they do not.

We present here a single example of a successful model using the EQ data. The remainder of the results related to the TTI Success Insights EQ questionnaire are presented in [172] and [173], which may be accessed from the TTI Success Insights Research webpage https://ttiresearch.com/.

Table 12.2: EQ Assemblers and Fabricators All Other
Logit Regression Model, N = 113

Variable	Estimate	Std. Error	z value	Pr(> |z|)	Sig†
(Intercept)	1.594	0.619	2.573	0.01	*
SOCIAL AWARENESS	-0.039	0.009	-4.343	0	***
AUC◇	0.643		Brier♭	0.188	

Note 12.1. † *variable significance may be interpreted as* . < 0.1, ∗ < 0.05, ∗∗ < 0.01, ∗∗∗ < 0.001.

Note 12.2. ⋄ *area under the curve, any score greater than* 0.50 *indicates the model is better than random sampling.*

Note 12.3. ♭ *Brier statistic is a goodness of fit measure for the receiver operating characteristic curve; the lower the score, the better the fit.*

The results presented in Table 12.2 show a reasonably solid model given that it is based on a completely randomly generated subset of the overall TTI Success Insights population and based solely on EQ data. As we have established in other studies, when we consider multiple sciences and have more uniquely defined data available, we generally have seen considerably stronger results. The area under the curve (AUC) statistic is acceptable if it is greater than 0.50, representing a model that performs better than a random selection model. Since the AUC in Table 12.2 exceeds 0.64, we are satisfied that this model performs better than a random selection model.

As an example, we present the results of a study conducted with a medical devices manufacturing company. The study was conducted on a set of 72 aortic valve sales personnel. We appended a random sample of just over 200 members from the general population of TTI Success Insights respondents. As can be seen, an AUC exceeding 0.80 is obtained when we have a more uniquely defined data set on multiple sciences. In this case, we have used our Talent Insights (comprised of our DISC and Motivator questionnaires) and our Acumen Capacities Index (our version of the Hartman Value Profile).

Table 12.3: Aortic Valve Sales Sample
Logit Regression Model, N = 72

Variable	Estimate	Std. Error	z value	Pr(> |z|)	Sig†
Intercept	4.95	1.6779	2.95	0.0031	**
People Oriented	-0.04	0.0133	-3.35	0.0008	***
Adapted C	-0.02	0.0073	-3.15	0.0016	**
Systems Judgment	-0.03	0.0107	-2.90	0.0037	**
Theoretical	-0.03	0.0088	-3.83	0.0001	***
Utilitarian	0.03	0.0085	3.67	0.0002	***
AUC⋄	0.82		Brier♭	0.12	

Note 12.4. † *variable significance may be interpreted as* . < 0.1, ∗ < 0.05, ∗∗ < 0.01, ∗∗∗ < 0.001.

Note 12.5. ⋄ *area under the curve, any score greater than* 0.50 *indicates the model is better than random sampling.*

Note 12.6. ♭ *Brier statistic is a goodness of fit measure for the receiver operating characteristic curve; the lower the score, the better the fit.*

Unfortunately, we do not have a large enough sample readily available with multiple sciences that includes our Emotional Quotient questionnaire as part of the data. Our future plans include conducting studies in this area to capture this information.

13 Response Processing Evidence of Validity

The American Psychological Association (APA) Handbook of Testing and Assessment in Psychology, see [75], states that assessment response processes require the collection of evidence demonstrating that the test taker is cognitively processing and properly interpreting the intended purpose of test items. This form of validity evidence is used to demonstrate that the assessment directs participants to engage in specific behaviors deemed necessary to complete the designed purpose of the assessment items.

Historically, response processing has been difficult to document. To further articulate the complexity of this task, we present the following quote from [163]. They state;

> Gathering validity evidence based on response process is perhaps the most difficult validity evidence to gather because it involves demonstrating that examinees are invoking the hypothesized constructs the test is designed to measure in responding to test items. As the Standards [4] describe, theoretical and empirical analyses of the response process of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees. Gathering this type of evidence is difficult because one cannot directly observe the cognitive processes going on within people's heads as they respond to test items. Although some studies have used MRI to see which regions of the brain are activated when responding to tasks, most studies of response processes use indirect means such as cognitive interviews, think-aloud protocols, focus groups, or analysis of answer patterns and item response time data and attempt to set the stage with introductory statements of purpose.

While the above quote suggests several approaches to expose cognitive processing, they all fail to provide the detail required to trace out actual mental decision-making processing. For example, magnetic resonance imaging (MRI) studies show regions of the brain that activate when responding to various stimuli, see [133], yet MRI only shows the resulting brain activity based on blood flow and is, therefore, unable to detect the moment-by-moment decision-making pathways leading to a respondent's cognitive processing.

However, during the design of a continual improvement process for all of TTI Success Insights' assessments, an opportunity was seen to combine the two authors' expertise to form a symbiotic relationship between state-of-the-art assessment analytics and neurophenomenological brain activity that captures real time decision-making pathways while participants are responding to assessment items. This unique mixed-methods approach mathematically identified weaker items and the brain activity exposes the mental processing that is helping TTI Success Insights understand the "why" and not just the "what".

13.1 Using Electroencephalograph (EEG) to Measure ACI Response Processing

Electroencephalography (EEG) measures voltage fluctuations within specific neural networks or regions of the brain. As a result, EEG records the brain's spontaneous electrical activity, thereby exposing the brain decision-making pathways. In June, 2015, TTI Success Insights was issued a patent for their work in Validation Processes for Ipsative Assessments, see [21] and [22]. The patent abstract reads:

> This invention is a validation process for ipsative assessments. Respondents are connected to an Electroencephalograph (EEG) and some or all of the ipsative assessment questions are asked again while connected to the EEG. The EEG measuring frontal lobe responses in terms of gamma waves is compared with the assessment questions. Positive responses proved one frontal lobe response in terms of gamma waves, negative or false answers provide a different gamma response, and neutral questions provide a neutral gamma response. Reading the responses then tells whether the respondent initially responded with integrity. If so, the assessment is validated.

Detailing all protocols used for our internal response validation process is beyond the scope of this manual. The interested reader may consult any of [25], [26], [27], [29], [30], [48], [49], or [53].

In general, the Gamma for Ipsative Validation using Electroencephalography (GIVE) process accesses asymmetric gamma wave bursts in the prefrontal cortex to validate the underlying subconscious decisions behind these self-report responses at the very moment of decision-making. As stated in the patent, until now no process has linked these specific types of self-reports to actual brain activity. The new process uses asymmetric wave analysis resulting from stimulus to validate the underlying mental decisions behind these reported responses, at the very moment of decision-making, thus exposing the true thoughts behind the responses and documenting potential abnormalities between their pre-assessments and their actual brain activity. This process provides evidence that an evoked emotionally laden response results in corresponding brain activity and documents both the intensity of human emotional responses and the directionality of the response.

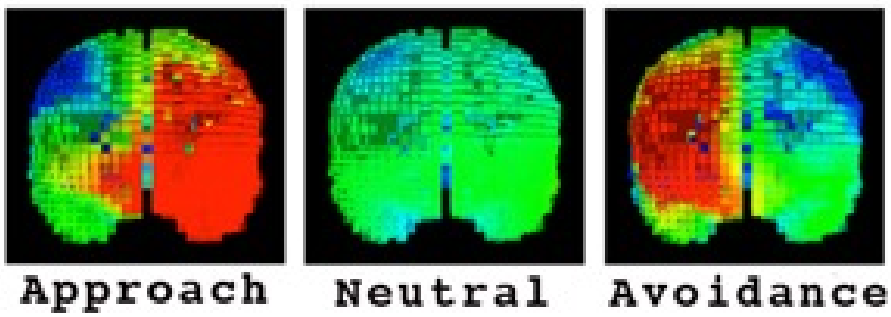


Figure 13.1: An Example of Asymmetry in EEG Captured Brain Activity

Figure 13.1 presents an example of frontal lobe gamma asymmetry with positive, neutral, and negative responses. The orientation of the brain is facing forward such that the right hemisphere is on the left side of the image. The red indicates an increase in gamma activity, the blue indicates a decrease in activity, and the green indicates little or no activation. In addition to color depicting a range of response intensity, the gamma burst location is also key to interpretation. A left frontal lobe flare is an indication of acceptance or a positive response and a right-side flare demonstrates avoidance or a negative response to a stimuli.

13.1.1 The Beginning

While the resulting TTI Success Insights response processing patents [21] and [22] and our present protocols, which examines prefrontal gamma asymmetry, continue to be used as we improve assess-

ment items, it is important to reflect on the actual path that led to this innovative item analysis procedure.

The beginnings can be traced to 1984 when Dr. Bonnstetter began to teach "The Science of Self" part of his science teacher preparation program. While the program gained national acclaim and was awarded the *Outstanding Secondary Science Education* in the nation, many in the the academic community were skeptical of using "self-reporting" instruments in an educational setting. The American Psychological Association (APA) Handbook of Testing and Assessment in Psychology states that the assessment response processes require the collection of evidence demonstrating that the test taker is cognitively processing and properly interpreting the intended purpose of the test items, see [163]. Early efforts by the APA involved asking others to verify the conclusions made by a participant, but no direct validation had been developed. Thus started TTI Success Insights' search for empirical, brain based evidence that could corroborate how a person responds to our assessment items.

Our first efforts can be traced back to an experiment involving the collection of electroencephalogram (EEG) data while participants were randomly shown each Hartman assessment item. Figures 13.2 and 13.3 show the result of this very first experiment. This initial EEG data processing involved an averaging of all bandwidths, so the imaging depicts an intensity of brain activity with the red showing greater brain activation. At this point, we were not collecting assessment responses and brain activity simultaneously. Therefore, the participant took the ACI assessment in the usual manner and then within a 72-hour timeframe, were shown each of the 18 items in a random order while we captured their brain reaction. Using standardized low-resolution brain electromagnetic tomography (sLORETA) to analyze and view voxel images of real-time brain activity, we were able to visualize participant decision-making processing. These images show average brain activity in the left hemisphere for each ACI item. At this point, we had not discovered the power of observing different brain regions or different brain activation frequencies. See the following references [23], [24], [25], [26], [29], [47], [50], [51], [52], and [54].

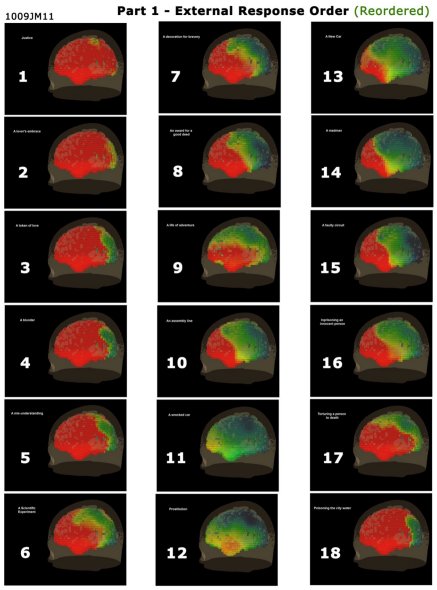


Figure 13.2: Confirmation Between the Survey and Brain Response
ACI World View

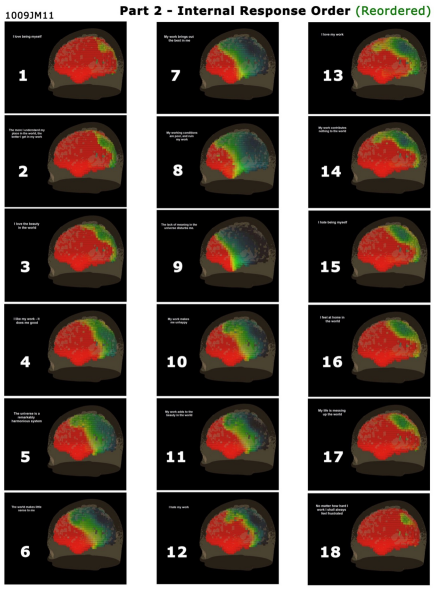


Figure 13.3: Confirmation Between the Survey and Brain Response
ACI Self View

13.2 Response Processing Summary

While response process evidence of validity continues to be a challenging assessment design hurdle for most assessment providers, the process of using sLORETA imaging from EEG data with QEEG (quantitative EEG) analysis and matching these insights to the population assessment analytics is opening new assessment item discussions that both affirm many mathematical findings and offer new possible explanations regarding respondent mental processing. We are entering a new era where industrial and organizational assessments can be revisited and refined to go beyond the simple approach of identifying mathematically or statistically weak items and expose the mental processing behind each of our decisions.

14 Evidence of Validity Based on Consequences of Testing

Validity evidence based on consequences of testing is summarized in the following definition:

Definition 14.1. *Validity evidence based on the consequences of testing refers to evaluating both the intended and unintended consequences associated with a testing program.*

It is clear a debate remains over whether testing consequences are "... extraneous to validity", see [163], or "... a critical element in evaluating the appropriateness of using a test for a particular purpose", see [127].

The APA appears to address the somewhat contradictory statements. The first claims that the consequences of testing are extraneous to establishing evidence of validity, while the second emphasizes evaluating a test's appropriateness for a specific purpose. As stated in [163]:

> We believe this debate is one of nomenclature, and given that virtually all testing programs have consequences on some level, it is important to evaluate the degree to which the positive outcomes of the test outweigh any negative consequences.

The takeaway appears to be that while evaluating the validity of testing consequences is essential, it remains unclear how much this aids in establishing an overall validity argument.

Based on the previous brief discussion, the primary question to consider is whether the test fulfills its intended purpose. To answer this, one must first determine the stated purpose of the test. Affirmative responses to this fulfillment question constitute evidence of positive consequences. Negative consequences are more complex and may include concepts such as adverse or disparate impact in hiring, education, and other contexts. Interpretation and use of the assessment also fall into this category, with an emphasis on identifying and mitigating any potential negative consequences associated with the assessment, its scoring, uses, and interpretation.

14.1 Additional Negative Consequences and Mitigation Efforts Acknowledged by TTI Success Insights

TTI Success Insights recognizes several potential negative consequences and systematically works to identify mitigation strategies for them:

1. Social Desirability Bias
    - Description: Respondents may answer or respond in was that they believe are more socially acceptable.
    - Consequence: The may lead to inaccurate results, with respondents underreporting undesirable traits or over-reporting desirable ones.
    - Mitigation: Our patented response processing, described in detail in Section 13, captures real-time brain images as test takers respond to items. This allows The TTI Success Insights research team to match responses to decision-making brain activity. Coupled with quantitative data, especially items indicating lower levels of difficulty and discrimination, this approach moves beyond identifying problematic items to understanding why they may need improvement.
2. Self-Insight Limitations

- Description: Individuals may lack insight into their own behavior, thoughts, or feelings, leading to incomplete or inaccurate responses.
- Consequence: This compromises the validity of self-reports, especially for complex traits or behaviors.
- Mitigation: TTI Success Insights addresses this concern by using a battery of assessment tools, not just a single assessment. For example, concerns identified in one tool may be cross-referenced with another. A strong behavioral style may gain additional explanation when compared to self-awareness and self-regulation levels within the Emotional Quotient Assessment.

3. Recall Bias

- Description: Self-report tests often rely on a person's memory, which can be flawed, especially for past events which occurred a long time ago.
- Consequence: This may lead to inaccuracies if respondents misremember past events or experiences.
- Mitigation: TTI Success Insights recognizes the crucial role of context in completing assessments and provides messaging to establish a focused mindset before individuals respond.

4. Response Style Bias

- Description: People may have consistent answering styles, such as always agreeing with statements (acquiescence bias) or avoiding extremes.
- Consequence: This can skew results, reducing the test's reliability and complicating the interpretation of true responses.
- Mitigation: Messaging reminds respondents that there are no right or wrong answers, encouraging honest responses for more accurate reporting.

5. Lack of Objectivity

- Description: Respondents' subjective perceptions may not align with objective reality.
- Consequence: Self-reports can differ from third-party observations or objective measures, potentially impacting reliability.
- Mitigation: 360-degree assessments can provide different perspectives, fostering richer communication and understanding.

6. Over-reliance on Self-perception

- Description: Self-report tests assume that respondents' self-perceptions align with actual traits or behaviors.
- Consequence: Distorted self-perceptions can skew data and lead to misunderstandings of underlying traits.
- Mitigation: TTI Success Insights' certification training establishes that self-reported assessments are just one of many tools that should be used in decision-making scenarios and should represent no more than one-third of the required information.

7. Emotional State Influence

    - Description: A respondents' current mood or emotions can influence responses.
    - Consequence: Temporary emotional states may impact responses, making results less stable over time.
    - Mitigation: Pre-assessment instructions highlight the need for participants to consider their emotional state before responding to the assessment. TTI Success Insights collects behind-the-scenes data, identifying possible "red flags" such as unusual timing or inconsistent response trends. Negative emotional influence is also evident when using multiple assessment tools and cross-referencing findings, such as changes between normal and adapted behavioral styles. Inconsistent results may be an opening for dialogue about these differences.

As an example, one popular use of the TTI Success Insights Emotional Quotient (EQ) assessment is in coaching and professional development. Generally speaking, TTI Success Insights itself does not directly participate in coaching or professional development; rather, a network of Value-Added Associates (VAA) fulfills this role. TTI Success Insights provides VAA with training and certification processes to ensure proper use and interpretation of the EQ assessment and its results.

15 Summary and Future Work

This manual is essentially comprised of four main areas. The first area is laying the general foundation for emotional intelligence (EI) by providing some theoretical background, an outline of the TTI Success Insights approach to Ei, and a more general discussion of the research philosophy at TTI Success Insights. The second section discusses the general areas of evidence of reliability, covering what we do and don't do as an organization. The third area presents our approach to evidence of validity, including some theoretical background on the various mathematical and statistical techniques employed. Finally, we provide a series of appendices containing our analyses of the various languages outside of North America for which we have versions of our EI assessment, along with some discussions of our approaches to test adaptation.

15.1 Summary of Introductory Sections

Our introductory sections begin with a discussion of the history of emotional intelligence (EI) and an outline of the TTI Success Insights approach to measuring EI, see Sections 2 and 3. This is then followed by a discussion on our thoughts related to the current state of measuring evidence of reliability and validity in Section 4. We include a discussion on the misunderstandings of the correlation versus causation argument. We discuss self-report bias and forced-rank versus likert-style response formats. Section 5 concludes the introductory sections with a brief summary of the American Psychological Association's guidance on measuring evidence of reliability and validity.

The aforementioned sections of this manual heavily cite external resources. The intent here is twofold. First, we want to ensure credit is given to those who have laid the groundwork over the last century and a quarter, and then some. Second, we wish to provide the interested reader with a comprehensive list of the resources we employ during our analyses. We hope that we have done an adequate job in both areas.

15.2 Summary of Evidence of Reliability Sections

Sections 4 and 5 really kick off the discussions of evidence of reliability and validity, but our approach to evidence of reliability is outlined in Sections 6 and 7. In Section 9 we discuss item response theory (IRT) and how certain outputs of the IRT modeling approach give us insight into a continuous view of reliability through the test information function (TIF). However, since IRT also provides information about evidence of internal structure validity, we postpone the presentation of any IRT results until Section 11.

In Sections 5.1.1 and 5.1.2, we discuss alternate forms reliability and generalizability theory and present our reasons for not measuring these areas of reliability. In the case of the former, it is simply cost prohibitive to have multiple versions of the same assessment measuring the same concepts for a comparison. Creating a reliable of valid psychometric assessment is difficult as it is, creating two different versions would be challenging at best. The cost of time, effort, and resources is prohibitive. In the case of the latter, we simply do not have data collection procedures in place to capture enough data and a granularity that makes generalizability theory a bona fide option. We have explored the area and we do like the outcomes we have seen, we just lack enough data to make this approach tenable at this time.

Section 6 presents our analyses based on internal consistency reliability. We present multiple versions of approaches to measuring this important area. Our results may be seen in Section 6.5. We first consider inter-item correlation analyses. In all cases, we see solid evidence of reliability

based on this approach. We next consider the quite popular and often cited α-coefficient. For the reader coming to the summary for the meat of this treatise, it is often and incorrectly assumed the Cronbach is the originator of the α-coefficient and it is often called Cronbach's alpha. As noted in Section 4.1, the actual originator is Kuder and Richardson in 1937, see [105]. Once again, all our constructs show very solid values of the α-coefficient with very tight confidence intervals.

A final analysis is presented in Section 6.5 known as the α-if-item-deleted statistic. This is an approach whereby an item is removed from consideration and the α-coefficient is then computed and compared to the value with the item considered. In this way, we measure the contribution to the internal consistency of the item. If the value of α increases significantly when an item is removed, one may question whether the item is adding any relevant information into the assessment. In all cases we see no items that are highlighted for questioning in this area.

Finally, we present our historical analytical results for internal consistency evidence of reliability in Section **??**. We present the historical record as part of our overall argument for internal consistency reliability. We have been tracking these type statistics for many years. However, the first version of this manual was produced in 2018. Earlier versions of this information exist and can be provided upon request.

The second area we measure from the APA's guidance is the so-called test-retest statistic. We prefer to use the term temporal consistency to complement the use of the term internal consistency and to highlight that we are not *testing* anyone, we are measuring their standing on a construct. However, the APA refers to it as test-retest. We are interested in ensuring that our population of respondents maintains a consistent standing relative to each other on successive iterations of responding to the questionnaire. The details of our approach are outlined in Section 7. We note here that due to our update of the questionnaire, we do not have sufficient data to adequately measure temporal consistency on the new questionnaire at this time. We do present our historical findings in Section 7.4.

15.3 Summary of Evidence of Validity Sections

We begin our discussion around evidence of validity in Section 8 where we outline our approach to construct definitions, construct relevance, construct representation, and test construction procedures. We also present an outline of the historical development of this questionnaire. We further present an outline of our assessment construction philosophy in Appendix R, as well as our assessment adaptation protocols in Appendix Q. The appendix on adaptations is really just a synopsis of the guidance on the topic provided by the International Test Commission. We did this to condense and summarize a host of highly technical information for the interested reader.

We next present our main mathematical analyses of the questionnaire and its respondent data in Section 11. To facilitate the discussion, we first present two sections on item response theory (IRT) and exploratory factor analysis (EFA), see Sections 9 and 10 respectively. We are measuring our evidence of internal structure validity in three ways. First, we consider the statistic (corrected) item-to-total-score correlation (CITC). As a primer, item-to-total-score correlation measures how much a given item is contributing to the score of a given construct. However, since the item is part of the score, to more adequately measure the item's contribution, we remove the item's responses from the scoring and then compute the correlation. Hence the addition of the term corrected into the name of the statistic.

Next, we consider IRT in which we compute a graded response model (GRM) for each of the constructs. GRM provides information on the difficulty and discrimination of each of the items,

as well as measuring both item level information and an overall test information statistic that are presented across the entire spectrum, at least in as much as is reasonable. What we mean is that the domain of the GRM models is the entire real line, i.e., from negative infinity all the way to positive infinity. However, most GRM functions are essentially zero outside of the range of negative 4 to positive 4.

Finally, we present our EFA results. We show how aligned each of the constructs is to the originator's theoretical outline as presented in Section 2 (EI History) and discussed in Section 8 (Test Content). We further present two statistics based on the output of the EFA models. The first, variance explained, is a measure of how much of the variation in the response patterns is due to random error versus being explained by the accuracy of the questionnaire items. The second is called composite reliability and is an additional measure of consistency, this time based on the EFA model as compared to the results obtained directly with the α-coefficient and those inferred from the IRT modeling approach.

We present results based on the current, updated questionnaire as well as our historical analyses, where applicable. The reason for the *where applicable* being appended to the end of the previous statement is that the original analyses for the first version of this model did not have an IRT approach available. We present as much information as we currently have available. Section 11.4 presents the current results, while Section B.1 presents our historical data.

Our analyses indicate that our evidence of validity based on internal structure is solid overall. We see a single item on the Social Awareness scale that struggles across all forms of the analysis. This item has been flagged and we are currently testing potential replacements as we are writing this manual. At such time as a suitable replacement is found, we will update the questionnaire and the analyses. All scales show solid CITC, excellent test information, and very solid variance explained and composite reliability. All scales show variance explained above 0.50, except for Social Awareness. The lack of exceeding 0.50 on variance explained is no doubt largely due to the aforementioned item that is going to be replaced. In falling short, this scale still comes in with a respectable 0.47.

We next present our approach to measuring evidence of validity based on relationships to external variables. Our approach is outlined in Section 12. However, we can only present historical results since we have recently made the decision to update the questionnaire in an attempt to increase the internal consistency and internal structure results. This means that we no longer have adequate numbers of data in the relevant areas to conduct an analysis of the data based on the new questionnaire. Depending on the velocity of our data collection with respect to this questionnaire, we hope to be able to add the analyses back into the next iteration of this questionnaire. We have historical data available upon request and have plans to release a *jobs manual* in the future that contains our historical results and will contain our future analyses when complete.

The last two areas of evidence of validity are somewhat interrelated. These are validity related to response processing and consequences of testing. The interrelatedness is due to the fact that we often use our neurological analyses conducted in our response processing research as an aide to our approaches to consequences of testing. We use the results of our research using EEG data collected during the active taking of our assessments to identify the good and the bad in our assessments, reports, training, etc. We then use this knowledge to improve, educate, and help develop mitigation strategies for areas we identify as needing improvement and so on. These areas of validity evidence are presented in Sections 13 and 14 respectively.

15.4 Summary of our Appendices

Appendices C through P contain the mathematical and statistical measures associated with measurement of evidence of reliability and validity as relates to the languages outside North American English. This includes 14 versions of the questionnaire in a total of 11 languages. We currently have three versions of English (North America to include the U.S. and Canada, Australia/New Zealand, and The United Kingdom), two versions of French (France and French Canadien), and two versions of Spanish (Spanish North America and Spanish-Spain).

It should be noted that there is likely a need for many more adaptations. For example, Spanish North America includes all Spanish speaking countries in North and South America that use our EQ assessment. It has been a business based decision to allow the use of this questionnaire throughout the western hemisphere, at least until such time as we are able to institute a full study in the various Spanish speaking nations using this tool. There is also likely a need for similar approaches for French, German, and English, at a minimum.

Even within North America alone, there may be a need for multiple versions of English for cultural reasons. For example, it may be more sensible to test whether there are differences in response patterns based on the use of our North American version versus, say, our UK version when sampling a population of English speaking Canadiens. The ITC is very specific about the need for an adaptation to take into account both the language and local culture. One may consult the ITC guidance on the subject to be found here. One such requirement is that the individuals performing the translation from the development language (North American English) be native speakers of the target language and reside in the geographic location where the adapted assessment is to be used. The second part of the requirement is an attempt to ensure an adaptation consistent with the culture where the adaptation is to be used.

A couple final notes on the language adaptations and their analyses is in order. Many of the languages have very low usage in a given year. We have provided our analysis for our distributors to have available with the caveat that the stability and reliability of any analyses are a function of the quantity and quality of the available data. In particular, with a new questionnaire for North American English getting set for full implementation at the time of the writing of this manual, all analyses presented for languages outside North American English are based on the previous questionnaire. At such time as each language has their translation completed, we will begin a testing process to see that the translated new questionnaire is a good fit for given language. If not, we will begin an iterative process designed to overcome any shortcomings brought on by cultural or linguistic issues.

The final two appendices contain a summary of the ITC guidance on translation and adaptation, see Appendix Q, and our questionnaire development process, see Appendix R.

15.5 Future Work

We have made many references to potential future projects throughout this manual. We attempt to itemize them here with some commentary, as appropriate.

1. Translation and Adaptation as needed.
2. Logistic regression models for US O*NET job classification system, see Section 12.
3. Other relevant studies designed to elicit evidence of validity based on relationships to other variables.

4. Continued research into potential replacement questionnaire items, as needed.

5. A follow on response process validation study for publication, see Section 13 as well as the protocol paper [29].

6. Other research studies designed to elicit information about the efficacy of our questionnaire to be used for identifying effective leadership, in particular at levels of Vice President or Senior Vice President and above within an organization (business, government, military, 501(c)-3, etc.).

7. Updated weighted scoring models as the data become available. The current iteration of scoring models is based on approximately 3,800 respondents.

8. Implement weighted scoring for adapted questionnaires, as appropriate.

9. Differential Item Functioning (DIF) studies as the data becomes available. We have done this in the past on the older questionnaire with some effectiveness. However, the new questionnaire requires we *start over*. The idea here is to identify any areas of the assessment where disparate impact may be introduced due to unforeseen bias inherent in the questionnaire or the underlying model itself.

References

[1] Caralee Adams. Report questions value of entrance exams in predicting college success. https://blogs.edweek.org/edweek/college_bound/2014/02/report_explores_use_of_standardize_test_scores_in_predicting_student_success.html, 2014. Education Week.

[2] Alan Agresti. *Categorical Data Analysis, Fifth Edition*. Pearson, New York, 2018.

[3] Neal M. Ashkanasy and Catherine S. Daus. Rumors of the death of emotional intelligence in organizational behavior are vastly exaggerated. *Journal of Organizational Behavior*, 26:441–452, 2005.

[4] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing, 2014.

[5] Elizabeth J. Austin, Ian J. Deary, Gavin J. Gibson, Murray J. McGregor, and J. Barry Dent. Individual response spread in self-report scales: Personality correlations and consequences. *Personality and Individual Differences*, 24(3):421–438, March 1998.

[6] James R. Averill and Elma P. Nunley. Grief as an emotion and as a disease: A social-constructionist perspective. In M.S. S troebe, W. Stroebe, and R.O. Hanson, editors, *Handbook of Bereavement: Theory, Research, and Intervention*. Cambridge University Press, New York, 1993.

[7] R. Michael Bagby, James D. A. Parker, and Grame J. Taylor. The twenty-item toronto alexithymia scale - i. item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, 38(1):23–32, January 1994.

[8] R. Michael Bagby, James D. A. Parker, and Grame J. Taylor. The twenty-item toronto alexithymia scale - ii. convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*, 38(1):33–40, January 1994.

[9] Frank B. Baker. *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, New York, 1992.

[10] Frank B. Baker and Seock-Ho Kim. *Item Response Theory Parameter Estimation Techniques*. Marcel Dekker, New York and Basel, 2nd, revised and extended, edition, 2004.

[11] J. D. Balakrishnan. Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance*, 25(5):1189–1206, 1999.

[12] Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. The Guilford Press, New York, 2018.

[13] Reuven Bar-On. *The Development of an Operational Concept of Psychological Well-being*. PhD thesis, Rhodes University, South Africa, 1988.

[14] Reuven Bar-On. Emotional quotient inventory: A measure of emotional intelligence. Technical report, Multi-Health Systems, Toronto, 1997.

[15] Reuven Bar-On. The emotional quotient inventory (EQ-i): Technical manual. Technical report, Multi-Health Systems, Toronto, 1997.

[16] W. L. Bedwell, S. M. Fiore, and E. Salas. Developing the 21st century (and beyond) workforce: A review of interpersonal skills and measurements strategies. https://atecentral.net/downloads/221/Salas_Firoe_IPS%20measurement%20Final%20draft.pdf, 2011. Paper prepared for the NRC Workshop on Assessing 21st Century Skills.

[17] A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, editors, *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.

[18] R.D. Bock. A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16(4):21–33, 1997.

[19] Kenneth A Bollen. *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics, 1989.

[20] Douglas G. Bonett and Thomas A. Wright. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, 2014.

[21] Bill J. Bonnstetter, Ronald J. Bonnstetter, Dustin Hebets, and Tom F. Collura. Validation process for ipsative assessments, 2016. US Patent 9,060,702.

[22] Bill J. Bonnstetter, Ronald J. Bonnstetter, Dustin Hebets, and Tom F. Collura. Validation process for ipsative assessments, 2016. Canadian Patent No. 2,808,691.

[23] R. J. Bonnstetter, D. Hebets, and N. L. Wigton. Frontal gamma asymmetry in response to soft skills stimuli: A pilot study. *NeuroRegulation*, 2(2):70–85, 2015. http://dx.doi.org/10.15540/nr.2.2.70.

[24] R.J. Bonnstetter. Review of the Book Altered Traits: Science Reveals How Meditation Changes Your Mind, Brain, and Body by D. Goleman and R.J. Davidson. *NeuroRegulation*, 5(3):103–104, 2018.

[25] Ronald J. Bonnstetter and Thomas F. Collura. Brain activation imaging in emotional decision making and mental health: A review – part 1. *Clinical EEG and Neuroscience*, 2020.

[26] Ronald J. Bonnstetter and Thomas F. Collura. Brain activation imaging in emotional decision making and mental health: A review – part 2. *Clinical EEG and Neuroscience*, 2020.

[27] Ronald J. Bonnstetter, Thomas F. Collura, and Dustin Hebets. Uncovering the Belief Behind the Action. *NeuroConnections*, 2012.

[28] Ronald J. Bonnstetter and Eric T. Gehrig. TTI Success Insights Style Insights 2020 Technical Manual Version 1.0. Technical report, TTI Success Insights, 2020. Forthcoming.

[29] Ronald J. Bonnstetter, Eric T. Gehrig, and Dustin Hebets. Response Process Validation Protocol using Neurophenomenological Gamma Asymmetry. *NeuroRegulation*, 2018.

[30] Ronald J. Bonnstetter, Dustin Hebets, and Nancy L. Wigton. Frontal gamma asymmetry in response to soft skills stimuli: A pilot study. *NeuroRegulation*, 2(2):70–85, 2020.

[31] Richard E. Boyatzis, Daniel Goleman, and Kenneth Rhee. Clustering competence in emotional intelligence: Insights from the emotional competence inventory (eci). In R. Bar-On and J.D.A. Parker, editors, *The Handbook of Emotional Intelligence: Theory, Development, Assessment, and Application at Home, School, and in the Workplace*, pages 343–362. Jossey-Bass: SanFrancisco, 2000.

[32] Travis Bradberry and Jean Greaves. *Emotional Intelligence Quick Book*. Simon and Schuster, New York, 2005.

[33] Anna Brown and Albert Maydeu-Olivares. Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurements*, 71(3):460–502, 2011.

[34] Anna Brown and Albert Maydeu-Olivares. Fitting a Thurstonian IRT model to forced-choice data. *Behavioral Research Methods*, 44:1135–1147, 2012.

[35] Anna Brown and Albert Maydeu-Olivares. How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18:36–52, 2013.

[36] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910.

[37] Ross Buck. *The Communication of Emotion*. The Guilford Press, New York, 1984.

[38] R.H. Burros. The estimation of the discriminal dispersion in the method of successive intervals. *Psychometrika*, 20:299–305, 1955.

[39] Li Cai. Factor Analysis of Tests and Items. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 5, pages 84–100. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.

[40] Salvatore J. Catanzaro and Jack Mearns. Measuring generalized expectancies for negative mood regulation: Initial scale development and implications. *Journal of Personality Assessment*, 54(3&4):546–563, 1990.

[41] J.M. Cattel. Mental tests and measurement. *Mind*, 15:373–380, 1890.

[42] Cary Cherniss and Daniel Goleman. The emotionally intelligent workplace: How to select for, measure, and improve emotional intelligence in individuals, groups, and organizations. In Cary Cherniss, Richard E. Boyatzis, and Maurice Elias, editors, *Advances in Emotional Intelligence: Research and Practice*. Jossey-Bass, 2001.

[43] Cary Cherniss and Daniel Goleman. *Guidelines for Best Practice*, 2014 (accessed May, 2018). http://www.eiconsortium.org.

[44] Eunseong Cho and Seonghoon Kim. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2):207–230, April 2015.

[45] Lee Anna Clark and David Watson. Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3):309–319, 1995.

[46] J. Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7:249–253, 1983.

[47] T. F. Collura, N. L. Wigton, C. Zalaquett, S. Chatters, and R.J. Bonnstetter. The Value of EEG-Based Electromagnetic Tomographic Analysis in Human Performance and Mental Health. *Biofeedback*, 2023.

[48] T. F. Collura, C. Zalaquett, and R. J. Bonnstetter. Seeing inside the client's mind. *Counseling Today*, 57(6):24–27, 2014.

[49] T. F. Collura, C. Zalaquett, R. J. Bonnstetter, and S. Chatters. Towards an operational model of decision making, emotional regulation, and mental health impact. *Advances in Mind-Body Medicine*, 28(4):18–33, 2014.

[50] T.F. Collura, R.J. Bonnstetter, D. Hebets, and B.J. Bonnstetter. Uncovering the Belief Behind the Action. *NeuroConnections*, Winter:20–23, 2012.

[51] T.F. Collura, R.J. Bonnstetter, and C.P. Zalaquett. Neurocounseling: Bridging Brain and Behavior. *Counseling Today*, 57(6):24–27, 2014.

[52] T.F. Collura, R.J. Bonnstetter, C.P. Zalaquett, and S.J. Chatters. Toward an Operational Model of Decision Making, Emotional Regulation, and Mental Health Impact. *Advances in Mind-Body Medicine*, 28(4):4–19, 2014.

[53] T.F. Collura, N.L. Wigton, C. Zalaquett, S. Chatters, and R.J. Bonnstetter. The Value of EEG-based Electromagnetic Tomographic Analysis in Human Performance and Mental Health. *Biofeedback*, 44(2), 2016.

[54] Thomas F. Collura, Nancy L. Wigton, Carlos Zalaquett, SeriaShia Chatters-Smith, and Ronald J. Bonnstetter. The Value of EEG-Based Electromagnetic Tomographic Analysis in Human Performance and Mental Health. *Biofeedback*, 44(2):58–65, 2016.

[55] International Test Commision. The ITC Guidelines for Translating and Adapting Tests (Second Edition). https://www.InTestCom.org, 2010.

[56] D.J. Cooke and C. Michie. An item response theory analysis of the hare psychopathy checklist-revised. *Psychological Assessment*, 9:3–14, 1997.

[57] Robert K. Cooper and Ayman Sawaf. *Executive EQ: Emotional Intelligence in Leadership and Organizations*. Grosset and Putnam, New York, 1990.

[58] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

[59] Lee J. Cronbach. My current thoughts on coefficient alpha and successor procedures. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education, University of California, Los Angeles, 2004.

[60] Lee J. Cronbach, Goldine C. Glesser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, Inc., New York, 1972.

[61] R.J. Davidson, D.C. Jackson, and N.H. Kalin. Emotion, plasticity, context, and regulation: Perspectives from affective neuroscience. *Psychological Bulletin*, 126(6), 2000.

[62] M. Davies, L. Stankov, and R.D. Roberts. Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology*, 75:989–1015, 1998.

[63] Robert F. DeVellis. *Scale Development: Theory and Applications*. Sage, Thousand Oaks, CA, 4th edition, 2017.

[64] William R. Dillon and Matthew Goldstein. *Multivariate Analysis: Methods and Applications*. Wiley Series in Probability and Mathematical Statistics, 1984.

[65] A.L. Edwards and L.L. Thurstone. An internal consistency check for scale values determined but the method of successive intervals. *Psychometrika*, 17:169–180, 1955.

[66] Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Psychology Press, Taylor & Francis Group, New York, NY, 2000.

[67] Robert J. Emmerling and Daniel Goleman. *Emotional Intelligence: Issues and Common Misunderstandings*, 2003 (accessed May, 2018). http://www.eiconsortium.org.

[68] Xitao Fan, Brent C. Miller, Kyung-Eun Park, Bryan W. Winward, Mathew Christiansen, Harold D. Grotevant, and Robert H. Tai. An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 18(3):223–244, 2006.

[69] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[70] Ronald A. Fisher. *Theory of Statistical Estimation*. John Wiley & Sons, New York, 1950. Proceedings of the Cambridge Philosophical Society 22 (1925) 700-725.

[71] W.P. Flannery and S.P. Reise adn K.F. Widaman. An item response theory analysis of the general and academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality*, 26:168–188, 1995.

[72] R.C. Fraley, N.G. Waller, and K.A. Brennan. An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78:350–365, 2000.

[73] Eric T. Gehrig. Classification of Serial Entrepreneurs via Logistic Regression: A Case Study. White Paper, October 2017.

[74] Kurt F. Geisinger. Reliability. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 2, pages 21–42. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.

[75] Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors. *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, 750 First Street NE, Washington, D.C. 20002-4242, 2013. American Psychological Association.

[76] Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors. *APA Handbook of Testing and Assessment in Psychology Volume 2: Testing and Assessment in Clinical and Counseling Psychology*, 750 First Street NE, Washington, D.C. 20002-4242, 2013. American Psychological Association.

[77] Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors. *APA Handbook of Testing and Assessment in Psychology Volume 3: Testing and Assessment in School Psychology and Education*, 750 First Street NE, Washington, D.C. 20002-4242, 2013. American Psychological Association.

[78] T. Giuliano and A. Swinkels. Development and Validation of the Mood Awareness Scale, 1992.

[79] Daniel Goleman. *Emotional Intelligence: Why it Can Matter More Than IQ*. Bantam Books, New York, 1995.

[80] Daniel Goleman. What makes a leader? *Harvard Business Review*, 76(131):93–104, 1998.

[81] Daniel Goleman. *Working with Emotional Intelligence*. Bantam Books (US) and Bloomsbury (UK), New York and London, 1998.

[82] Daniel Goleman. *An EI-based Theory of Performance*, 2004 (accessed May, 2018). http://www.eiconsortium.org.

[83] Daniel Goleman. *Emotional Intelligence: Issues in Paradigm Building*, 2004 (accessed May, 2018). http://www.eiconsortium.org.

[84] Daniel Goleman. *Social Intelligence: The New Science of Human Relationships*. Bantam Dell: A Division of Random House, Inc., New York, 2006.

[85] Daniel Goleman. *Leadership: The Power of Emotional Intelligence*. More Than Sound, Northhampton, MA, 2011.

[86] Daniel Goleman. *The Brain and Emotional Intelligence: New Insights*. More Than Sound, Northhampton, MA, 2011.

[87] Daniel Goleman. *They've Taken Emotional Intelligence Too Far*, 2011 (accessed May, 2018). http://ideas.time.com/2011/11/01/theyve-taken-emotional-intelligence-too-far/.

[88] Daniel Goleman, Richard E. Boyatzis, and Annie McKee. *Primal Leadership: Realizing the Power of Emotional Intelligence*. Harvard University Press, Boston, 2002.

[89] Richard L. Gorsuch. *Factor Analysis*. Lawrence Erlbaum Associates, 2nd edition, 1983.

[90] Bert F. Green. In defense of measurement. *American Psychologist*, 33:664–670, 1978.

[91] Mark T. Greenberg, Roger P. Weissberg, Mary Utne O'Brien, Joseph E. Zins, Linda Fredericks, Hank Resnik, and Maurice J. Elias. Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist*, 58(6-7):466–474, 2003.

[92] James W. Grice. Computing and evaluating factor scores. *Psychological Methods*, 6(4):430–450, 2001.

[93] H. Gulliksen. *Theory of Mental Test Scores*. Wiley, New York, 1950.

[94] Ronald K. Hambleton, Wim J. van der Linden, and Craig S. Wells. IRT Models for the Analysis of Polytomously Scored Data: Brief and Selected History of Model Building Advances. In Micheal L. Nering and Remo Ostini, editors, *Handbook of Polytomous Item Response Theory* , chapter 2, pages 21–42. Routledge: Taylor & Francis Group, 2010.

[95] Harry H. Harman. *Modern Factor Analysis*. University of Chicago Press, Chicago and London, third revised edition edition, 1976.

[96] Rink Hoekstra, Jorien Vugteveen, Peter M. Kruyen, and Matthijs J. Warrens. An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology*, 22(4):351–364, 2019.

[97] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, Inc., 2000.

[98] Scott Jaschik. High school grades: Higher and higher. https://www.insidehighered.com/admissions/article/2017/07/17/study-finds-notable-increase-grades-high-schools-nationally, 2017. Inside Higher Ed.

[99] Charles E. Johnson, Robert Wood, and S.F. Blinkhorn. Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61:153–162, 1988.

[100] Peter J. Jordan, Neal M. Ashkanasy, Charmine E.J. Härtel, and Gregory S. Hooper. Workshop emotional intelligence scale development and relationship to team process effectiveness and goal focus. *Human Resources Management Review*, 12(2):195–214, 2002.

[101] Robert M. Kaplan and Dennis P. Saccuzzo. *Psychological Testing: Principles, Applications, and Issues*. Wadsworth, Belmont, CA, 2001.

[102] Jae-On Kim and Charles W. Mueller. Factor analysis: Statistical methods and practical issues. Quantitative Applications in the Social Sciences, Sage Publications, Beverly Hills and London, 1982.

[103] Jae-On Kim and Charles W. Mueller. Introduction to factor analysis: What it is and how to do it. Quantitative Applications in the Social Sciences, Sage Publications, Beverly Hills and London, 1982.

[104] Rex B. Kline. *Principles and Practice of Structural Equation Modeling*. The Guilford Press, 4th edition, 2016.

[105] G. F. Kuder and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937.

[106] Richard D. Lane, Lee Sechrest, Robert Reidel, Victoria Weldon, Alfred Kaszniak, and Gary E. Schwartz. Impaired verbal and nonverbal emotion recognition in alexithynia. *Psychosomatic Medicine*, 58:203–210, 1996.

[107] D.N. Lawley. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, Series A*, 23:273–287, 1943.

[108] P.F. Lazarsfeld. The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen, editors, *Measurement and Prediction*. Princeton, NJ: Princeton University Press, 1950.

[109] Jane Loevinger. The attenuation paradox in test theory. *Psychological Bulletin*, 51:493–504, 1954.

[110] F. M. Lord. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 18:57–76, 1952.

[111] F. M. Lord. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13:517–548, 1952.

[112] F. M. Lord. A theory of test scores. *Psychometric Monograph, (No. 7)*, 1952.

[113] Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale, NJ, 1980.

[114] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Reading, MA, 1968.

[115] Richard E. Lucas and Brendan M. Baird. Global self-assessment. In Michael Eid and Ed Diener, editors, *Handbook of Mutlimethod Measurement in Psychology*, pages 29–42. American Psychological Association, Washington, DC, 2006.

[116] Albert Maydeu-Olivares and Ulf Böckeenholt. Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, 10(3):285–304, 2009.

[117] Albert Maydeu-Olivares and Anna Brown. Ordinal factor analysis of graded preference questionnaire data. *Multivariate Behavioral Research*, 45:935–974, 2010.

[118] John D. Mayer and Deborah Bremer. Assessing mood with affect-senstive tasks. *Journal of Personality Assessment*, 49:95–99, 1985.

[119] John D. Mayer, David R. Caruso, and Peter Slovey. Emotional intelligence meets traditional standards for and intelligence. *Intelligence*, 27(4):267–298, 1999.

[120] John D. Mayer, Maria DiPaolo, and Peter Salovey. Perceiving affective content in ambiguous visual stimuli: A component of emotional intelligence. *Journal of Personality Assessment*, 54:772–781, 1990.

[121] John D. Mayer and Glenn Geher. Emotional intelligence and the identification of emotion. *Intelligence*, 22:89–113, 1996.

[122] John D. Mayer and Peter Salovey. What is emotional intelligence. In P. Salovey and D.J. Sluyter, editors, *Emotional Development and Emotional Intelligence: Educational Implications*, pages 3–34. Harper Collins: New York, 1997.

[123] John D. Mayer, Peter Salovey, and David R. Caruso. Models of emotional intelligence. In R. Sternberg, editor, *Handbook of Intelligence*, pages 396–420. Cambridge University Press, 2000.

[124] John D. Mayer and Alexander A. Stevens. An emerging understanding of the reflective (meta-) experience of mood. *Journal of Research in Personality*, 28:351:373, 1994.

[125] Roderick P. McDonald. *Factor Analysis and Related Methods*. Lawrence Erlbaum Associates, 1985.

[126] Roderick P. McDonald. *Test Theory: A Unified Treatment*. Taylor and Francis, 1999.

[127] Samuel Messik. Meaning and values in test evaluation: The science of ethics and assessment. *Educational Researcher*, 18(2):5–1, 1989.

[128] Orval H. Mowrer. *Learning Theory and Symbolic Processes*. Wiley, New York, 1960.

[129] Michael L. Nering and Remo Ostini, editors. *Handbook of Polytomous Item Response Theory Models*. Routledge: Taylor & Francis Group, New York and London, 2010.

[130] Jum C. Nunnally. *Psychometric Theory*. McGraw-Hill, New York, 2nd edition, 1978.

[131] Thomas Oakland, Ype H. Poortinga, Justin Schlegal, and Ronald K. Hambleton. International test commission: Its history, current status, and future directions. *International Journal of Testing*, 1(1):3–32, 2001.

[132] Remo Ostini and Michael L. Nering. Polytomous Item Response Theory Models. In *Quantitative Applications in the Social Sciences*, volume 144. Sage Publications, 2006.

[133] W.J. Owen, R. Borowsky, and G.E. Sarty. Fmri of two measures of phonological processing in visual word recognitionl. ecological validity matters. *Brain and Language*, 90:40–46, 2004. doi:10.1016/S0093-934X(03)00418-8.

[134] Tibor P. Palfai and Peter Salovey. The influence of depressed and elated mood on deductive and inductive reasoning. *Imagination, Cognition, and Personality*, 13:57–71, 1993.

[135] Benjamin R. Palmer. Models and measures of emotional intelligence. *Organisations and Peoiple*, 14(2):3–10, May 2007.

[136] Delroy L. Paulhus and Simine Vazire. The self-report method. In Richard W. Robins, Chris R. Fraley, and Robert F. Krueger, editors, *Handbook of Research Methods in Personality Psychology*, pages 224–239. Guildford, New York, 2007.

[137] W. L. Payne. *A Study of Emotion: Developing Emotional Intelligence; Self-integration; Relating to Fear, Pain, and Desire*. PhD thesis, Union Graduate School, 1983.

[138] Randall D. Penfield. Item Analysis. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 7, pages 121–138. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.

[139] Gjalt-Jorn Y. Peters. The alpha and omega of scale reliability and validity. *The European Health Psychologist*, 16(2):56–69, April 2016.

[140] David M. W. Powers. Evaluation: From Precision Recall, and F-Factor to ROC, Informedness, Markedness, and Correlation. Technical report, Flinders University of South Australia, School of Informatics and Engineering, December 2007.

[141] G. Rasch. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: The Danish Institute for Educational Research, 1960. G. Leunbach, Trans.

[142] G. Rasch. An individualistic approach to item analysis. In P.F. Lazarsfeld and N.W. Henry, editors, *Readings in Mathematical Social Science*, pages 89–107. Cambridge: MIT Press, 1966.

[143] G. Rasch. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14:58–94, 1977.

[144] Tenko Raykov and George A. Marcoulides. *Introduction to Psychometric Theory*. Routledge, New York, 2011.

[145] Steven P. Reise, Tyler M. Moore, and Mark G. Haviland. Applying Unidimensional Item Response Theory Models to Psychological Data. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 6, pages 101–119. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.

[146] William Revelle and Richard E. Zinbarg. Coefficients alpha, beta, omega, and the glb: comments on sijtsma. *Psychometrika*, 74(1):145–154, 2009.

[147] H.J. Rimoldi and M. Hormaeche. The law of comparative judgment in the successive intervals and graphic rating scale methods. *Psychometrika*, 20:307–318, 1955.

[148] Robert Rosenthal, Judith A. Hall, M. Robin DiMatteo, Peter L. Rogers, and Dane Archer. *Sensitivity to Nonverbal Communication: The PONS Test*. Johns Hopkins University Press, New York, 1997.

[149] Peter Salovey and Deborah Birmbaum. Influence of mood on health-relevant cognitions. *Journal of Personality and Social Psychology*, 57(3):539–551, 1989.

[150] Peter Salovey and John D. Mayer. Emotional intelligence. *Imagination, Cognition, and Personality*, 9(3):185–211, 1990.

[151] Peter Salovey, John D. Mayer, and David R. Caruso. The positive psychology of emotional intelligence. In C. R. Snyder and S.J. Lopez, editors, *Handbook of Positive Psychology*. Oxford University Press, Inc., New York, 2001.

[152] Peter Salovey, John D. Mayer, Susan Lee Goldman, Carolyn Turvey, and Tibor P. Palfai. Emotional attention, clarity, and repair: Exploring emotional intelligence using the trait meta-mood scale. In J. W. Pennebaker, editor, *Emotion, Disclosure, and Health*, pages 125–154. American Psychological Association, 1995.

[153] Peter Salovey and David J. Sluyter. *Emotional Development and Emotional Intelligence: Educational Implications*. Basic Books, A Division of Harper-Collins Publishers, New York, 1997.

[154] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17, 1969.

[155] F. Samejima. A general model for free response data. *Psychometrika Monograph Supplement*, 18, 1972.

[156] F. Samejima. The grade response model of the latent trait theory and tailored testing. In C.L. Clark, editor, *Proceedings of the First Conference on Computerized Adaptive Testing*, pages 5–17. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1976. (Professional Series 75-6).

[157] F. Samejima. Constant information model on the dichotomous response level. In D.J. Weiss, editor, *The 1979 Conference on Computerized Adaptive Testing*, pages 145–165. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1979.

[158] F. Samejima. Evaluation of the mathematical models for ordered polychotomous responses. *Behvariometrika*, 23:17–35, 1996.

[159] F. Samejima. Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, 62:471–493, 1997.

[160] Nicola S. Schutte, John M. Malouff, Lena E. Hall, Donald J. Haggerty, Joan T. Cooper, Charles J. Golden, and Liane Dornheim. Development and validation of a measure of emotional intelligence. *Personality and Individual Differences*, 25:167–177, 1998.

[161] Nicole Shectman, Louise Yarnall, Regie Stites, and Britte Cheng. Empowering adults to thrive at work: Personal success skills for 21st century jobs, a report on promising research and practice. https://www.sri.com/sites/default/files/publications/joyceempoweringadultstothriveatwork_4.pdf, 2016. The Joyce Foundation, Chicago, IL.

[162] S.G. Sireci, L Patsula, and R.K. Hambleton. Statistical Methods for Identifying Flaws in the Test Adaptation Process. In R.K. Hambleton, P. Merenda, and C. Spielberger, editors, *Adapting Educational and Psychological Tests for Cross-cultural Assessment*, pages 93–116. Lawrence Erlbaum Publishers, Mahwah, NJ, 2005.

[163] Stephan G. Sireci and Tia Sukin. Test Validity. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 4, pages 61–84. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.

[164] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[165] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910.

[166] Michael Spivak. *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Addison-Wesley Publishing Company, 1965.

[167] Jon Starkweather. Step out of the past: Stop using coefficient alpha; there are better ways to calculate reliability. *Research and Statistical Support, University of North Texas*, 2012.

[168] Alan Swinkels and Traci A. Giuliano. The measurement and conceptualization of mood awareness: Monitoring and labeling one's mood states. *Personality and Social Psychology Bulletin*, 21:934–939, 1995.

[169] Wei Tang, Ying Cui, and Oksana Babenko. Internal consistency: Do we really know what it is and how to assess it? *Journal of Psychology and Behavioral Science*, 2(2):205–220, June 2014.

[170] L. L. Thurstone. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16:433–451, 1925.

[171] L. L. Thurstone. *Multiple Factor Analysis*. University of Chicago Press, Chicago, 1947.

[172] TTI Success Insights, Ltd., Scottsdale, AZ. *TTI Success Insights Emotional Quotient Evidence of Validity Based on Relationships to External Variables*, 1 edition, August 2024.

[173] TTI Success Insights, Ltd., Scottsdale, AZ. *TTI Success Insights Evidence of Validity Based on Relationships to External Variables*, 1 edition, August 2024.

[174] L. R. Tucker. Maximum validity of a test with equivalent items. *Psychometrika*, 11:1–13, 1952.

[175] A.Y. Wang, R.P. Tett, R. Fisher, J.G. Griebler, and A.M. Martinez. Testing a model of emotional intelligence. Unpublished Manuscript, University of Central Florida, 1997.

[176] D.J. Weiss. Introduction. In D.J. Weiss, editor, *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, chapter 1, pages 1–8. New York: Academic Press, 1983.

[177] Rand R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing: A Volume in Statistical Modeling and Decision Science*. Academic Press, New York, fourth edition edition, 2017.

[178] Edward W. Wiley, Noreen M. Webb, and Richard J. Shavelson. The Generalizability of Test Scores. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 3, pages 43–60. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.

[179] Richard E. Zinbarg, William Revelle, Iftah Yovel, and Wen Li. Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1):123–133, March 2005.

[180] Joseph E. Zins, Roger P. Weissberg, Margaret C. Wang, and Herbert J. Walberg. *Building Academic Success on Social and Emotional Learning: What Does the Research Say?* Teachers College Press, New York, 2004.

A Historical Reliability Analysis

A.1 General Population Results

A.1.1 2025 General Population Results

Table A.1: U.S. General Population Statistics Standard Scoring
N = 4,395

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	78.24	11.39	77.35	11.71	79.14	11.00
Self Regulation	64.35	14.01	66.56	13.47	62.13	14.19
Motivation	74.53	12.23	75.66	11.73	73.42	12.60
Social Awareness	72.10	11.79	69.13	12.19	75.06	10.57
Social Regulation	72.70	12.40	72.47	12.51	72.90	12.22

Table A.2: U.S. General Population Statistics Weighted Scoring
N = 4,395

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	78.08	11.49	77.11	11.82	79.05	11.07
Self Regulation	64.84	13.95	67.08	13.40	62.60	14.14
Motivation	75.40	12.00	75.56	11.43	73.23	12.44
Social Awareness	72.44	11.89	69.51	12.30	75.36	10.69
Social Regulation	72.72	12.33	72.53	12.44	72.90	12.22

Table A.3: U.S. General Population Statistics
N = 3,803

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	78.64	11.76	77.61	11.78	79.66	11.66
Self Regulation	64.76	14.42	67.08	13.47	62.44	14.97
Motivation	74.61	12.76	75.68	12.40	73.54	13.02
Social Awareness	73.64	12.56	70.32	13.06	76.96	11.08
Social Regulation	72.55	13.11	72.73	13.10	72.37	13.12

It is worth taking the time to make note of the relatively significant drop in the Self Regulation mean scores going from the previous version of the EQ questionnaire (Table A.4) to the newer version of the questionnaire (Table A.3). Generally speaking, psychometric assessment scoring is

TTI SUCCESS INSIGHTS®

a measure of some underlying quantity that is assumed to be normally distributed. In particular, the underlying scoring methodology assumes a standard normal distribution, that is, a normal distribution with mean 0 and standard deviation 1.

A.1.2 2023 General Population Results

Table A.4: U.S. General Population Statistics 2023
N = 87,704

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	76.89	10.40	75.50	10.52	78.28	10.09
Self Regulation	70.90	12.14	72.27	11.87	69.53	12.26
Motivation	76.93	11.15	77.65	11.05	76.21	11.20
Social Awareness	72.56	11.61	69.86	11.88	75.27	10.66
Social Regulation	71.29	11.57	71.02	11.78	71.56	11.34

Extrapolating this assumption out somewhat and providing more clear-cut interpretation leads to the understanding that any normal distribution is defined on the entire real line. Moreover, we then map these scores to scores that are more friendly to the casual observer, in this case on a scale of 0-100. A standard normal distribution based score should then have a mean of 50 with a relevant standard deviation defined by the variation in the scoring data. The further above or below 50 a given scale's mean happens to be is an indication of how well balanced the population's responses are. The decrease in mean scoring on the Self Regulation scale is indicating a more well-balanced scale in comparison to some others.

A.2 U.S. Historical Inter-Item Correlation Analysis

This section presents the historical[4] results of the inter-item correlation study for the English-U.S. version. The results are presented in Table A.7. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

A.2.1 2025 IIC Results

Table A.5: Average U.S. Inter-item Correlations Weighted 2025

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.46	0.35	0.38	0.27	0.38
Item 2	0.43	0.36	0.38	0.27	0.37
Item 3	0.42	0.37	0.38	0.27	0.36
Item 4	0.44	0.36	0.37	0.26	0.35
Item 5	0.41	0.35	0.37	0.27	0.38
Item 6	0.45	0.37	0.37	0.29	0.37
Item 7	0.41	0.36	0.38	0.30	0.36
Item 8	0.44	0.35	0.38	0.28	0.37
Item 9	0.41	0.35	0.37	0.27	0.36
Item 10	0.43	0.35	0.37	0.25	0.36
Item 11		0.36	0.37	0.25	0.37
Item 12		0.35	0.37	0.38	
Mean	0.43	0.36	0.37	0.27	0.37
StDev	0.02	0.01	0.00	0.01	0.01

Table A.6: Average U.S. Inter-item Correlations 2025 Standard

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.47	0.36	0.39	0.29	0.41
Item 2	0.44	0.37	0.39	0.29	0.40
Item 3	0.43	0.38	0.39	0.29	0.38
Item 4	0.45	0.37	0.38	0.29	0.38
Item 5	0.42	0.36	0.39	0.29	0.40
Item 6	0.45	0.38	0.39	0.31	0.39
Item 7	0.42	0.36	0.40	0.33	0.38
Item 8	0.45	0.36	0.39	0.30	0.39
Item 9	0.42	0.36	0.39	0.30	0.39
Item 10	0.44	0.36	0.39	0.27	0.38
Item 11		0.36	0.39	0.27	0.39
Item 12		0.35	0.38	0.30	
Mean	0.44	0.36	0.39	0.29	0.39
StDev	0.02	0.01	0.00	0.01	0.01

[4]TTI Success Insights added the standard deviation of the inter-item correlations in 2023. Prior years included only the mean of the correlations.

A.2.2 2023 IIC Results

Table A.7: Average U.S. Inter-item Correlations 2023

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.31	0.27	0.28	0.25	0.26
Item 2	0.32	0.28	0.27	0.24	0.28
Item 3	0.30	0.29	0.28	0.25	0.27
Item 4	0.29	0.27	0.30	0.25	0.28
Item 5	0.28	0.28	0.28	0.24	0.26
Item 6	0.30	0.28	0.29	0.26	0.26
Item 7	0.30	0.29	0.27	0.27	0.26
Item 8	0.28	0.28	0.27	0.25	0.25
Item 9	0.30	0.28	0.26	0.24	0.26
Item 10	0.30	0.29	0.27	0.26	0.26
Item 11		0.27	0.28	0.23	0.29
Item 12		0.29	0.27	0.23	
Mean	0.30	0.28	0.28	0.25	0.27
StDev	0.01	0.01	0.01	0.01	0.01

A.2.3 2020 IIC Results

Table A.8: Average U.S. Inter-item Correlations 2020

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.23	0.35	0.24	0.23	0.30
Item 2	0.24	0.29	0.31	0.30	0.20
Item 3	0.30	0.25	0.25	0.23	0.26
Item 4	0.32	0.33	0.16	0.21	0.23
Item 5	0.36	0.27	0.24	0.39	0.29
Item 6	0.29	0.27	0.21	0.21	0.28
Item 7	0.27	0.22	0.29	0.14	0.30
Item 8	0.36	0.29	0.30	0.23	0.35
Item 9	0.29	0.31	0.34	0.27	0.30
Item 10	0.28	0.21	0.33	0.21	0.29
Item 11		0.30	0.28	0.33	0.18
Item 12		0.23	0.30	0.33	
Mean	0.29	0.28	0.27	0.25	0.27

A.2.4 2018 IIC Results

Table A.9: Average U.S. Inter-item Correlations 2018

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.23	0.35	0.24	0.23	0.30
Item 2	0.24	0.29	0.31	0.30	0.20
Item 3	0.30	0.25	0.25	0.23	0.26
Item 4	0.32	0.33	0.16	0.21	0.23
Item 5	0.36	0.27	0.24	0.39	0.29
Item 6	0.29	0.27	0.21	0.21	0.28
Item 7	0.27	0.22	0.29	0.14	0.30
Item 8	0.36	0.29	0.30	0.23	0.35
Item 9	0.29	0.31	0.34	0.27	0.30
Item 10	0.28	0.21	0.33	0.21	0.29
Item 11		0.30	0.28	0.33	0.18
Item 12		0.23	0.30	0.33	
Mean	0.29	0.28	0.27	0.25	0.27

A.3 U.S. Historical Internal Consistency via the α Coefficient

This section presents the historical[5] results of the internal consistency study for the English-US version. The results are presented in Tables A.12 through A.14. Table 6.1 may be consulted for the relevant scoring categories.

A.3.1 2025 Internal Consistency Results

Table A.10: Average U.S. α Coefficient Data 2025 Weighted

Est./Scale	SA	SR	MO	SoA	SoR
α	0.88	0.86	0.87	0.81	0.86
Lower CI	0.88	0.86	0.87	0.80	0.86
Upper CI	0.89	0.87	0.88	0.82	0.87

Table A.11: Average U.S. α Coefficient Data 2025 Standard

Est./Scale	SA	SR	MO	SoA	SoR
α	0.89	0.87	0.88	0.83	0.87
Lower CI	0.88	0.87	0.87	0.82	0.87
Upper CI	0.89	0.89	0.89	0.84	0.88

A.3.2 2023 Internal Consistency Results

Table A.12: Average U.S. α Coefficient Data 2023

Est./Scale	SA	SR	MO	SoA	SoR
α	0.80	0.82	0.80	0.79	0.79
Lower CI	0.80	0.82	0.80	0.79	0.78
Upper CI	0.80	0.82	0.81	0.80	0.79

[5]TTI Success Insights incorporated the α-if-item-deleted analysis in our 2023 reliability study. Prior years only included α coefficients with confidence intervals.

A.3.3 2020 Internal Consistency Results

Table A.13: Average U.S. α Coefficient Data 2020

Est./Scale	SA	SR	MO	SoA	SoR
α	0.79	0.82	0.79	0.82	0.80
Lower CI	0.76	0.81	0.76	0.80	0.78
Upper CI	0.81	0.82	0.80	0.83	0.82

A.3.4 2018 Internal Consistency Results

Table A.14: Average U.S. α Coefficient Data 2018

Est./Scale	SA	SR	MO	SoA	SoR
α	0.79	0.82	0.79	0.82	0.80
Lower CI	0.76	0.81	0.76	0.80	0.78
Upper CI	0.81	0.82	0.80	0.83	0.82

A.4 U.S. Historical Internal Consistency via the α-if-item-deleted

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

A.4.1 2025 α-if-item-deleted

Table A.15: α-if-item-deleted for U.S. Weighted 2025

Est./Scale	SA	SR	MO	SoA	SoR
α	0.88	0.87	0.88	0.82	0.86
Item 1	0.88	0.85	0.86	0.79	0.86
Item 2	0.87	0.85	0.87	0.79	0.85
Item 3	0.87	0.86	0.87	0.80	0.85
Item 4	0.88	0.85	0.86	0.79	0.84
Item 5	0.86	0.85	0.86	0.79	0.86
Item 6	0.88	0.86	0.86	0.81	0.85
Item 7	0.86	0.86	0.87	0.82	0.85
Item 8	0.88	0.85	0.87	0.80	0.85
Item 9	0.86	0.85	0.86	0.80	0.85
Item 10	0.87	0.85	0.86	0.78	0.84
Item 11		0.85	0.86	0.78	0.85
Item 12		0.85	0.86	0.80	

Table A.16: α-if-item-deleted for U.S. 2025

Est./Scale	SA	SR	MO	SoA	SoR
α	0.89	0.87	0.88	0.83	0.88
Item 1	0.89	0.85	0.87	0.81	0.87
Item 2	0.88	0.86	0.87	0.81	0.87
Item 3	0.87	0.87	0.87	0.81	0.86
Item 4	0.88	0.86	0.86	0.80	0.86
Item 5	0.87	0.85	0.87	0.81	0.87
Item 6	0.88	0.86	0.87	0.82	0.86
Item 7	0.87	0.86	0.88	0.83	0.86
Item 8	0.88	0.85	0.87	0.82	0.86
Item 9	0.87	0.85	0.87	0.82	0.86
Item 10	0.88	0.85	0.87	0.80	0.86
Item 11		0.85	0.87	0.80	0.86
Item 12		0.85	0.87	0.82	

A.4.2 2023 α-if-item-deleted

Table A.17: α-if-item-deleted for U.S. 2023

Est./Scale	SA	SR	MO	SoA	SoR
α	0.80	0.82	0.80	0.79	0.79
Item 1	0.81	0.80	0.81	0.79	0.78
Item 2	0.81	0.81	0.80	0.77	0.80
Item 3	0.79	0.82	0.81	0.79	0.79
Item 4	0.79	0.80	0.82	0.79	0.79
Item 5	0.78	0.81	0.81	0.77	0.78
Item 6	0.80	0.81	0.82	0.79	0.78
Item 7	0.80	0.82	0.80	0.80	0.78
Item 8	0.78	0.81	0.80	0.79	0.77
Item 9	0.79	0.81	0.80	0.78	0.78
Item 10	0.80	0.82	0.80	0.79	0.78
Item 11		0.81	0.81	0.77	0.80
Item 12		0.82	0.80	0.77	

B Historical Internal Structure Analysis

B.1 Historical Results

This section contains a historical record of the analyses related to evidence of validity based on internal structure. It is the intent of these authors to keep an ongoing, historical record that may be viewed by anyone reading this manual. The information provided is intended to aid the establishment of an argument for the validity of the TTI Success Insights Emotional Quotient assesssment.

B.1.1 US Corrected Item Total Correlation Analysis 2025

This section presents the results of the item-total correlation study for the English-US version. The results are presented in Table B.1. As is discussed in Section 11 above, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table B.1: US Corrected Item-total Correlations 2025 Current

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.40	0.61	0.56	0.48	0.43
Item 2	0.61	0.52	0.54	0.52	0.51
Item 3	0.71	0.45	0.51	0.47	0.60
Item 4	0.53	0.51	0.66	0.56	0.66
Item 5	0.72	0.61	0.58	0.49	0.45
Item 6	0.51	0.46	0.57	0.37	0.56
Item 7	0.72	0.51	0.49	0.18	0.63
Item 8	0.55	0.59	0.53	0.45	0.54
Item 9	0.74	0.58	0.57	0.46	0.59
Item 10	0.62	0.63	0.55	0.66	0.64
Item 11		0.56	0.57	0.65	0.55
Item 12		0.59	0.63	0.39	
Mean	0.61	0.55	0.56	0.47	0.56
StDev	0.11	0.06	0.05	0.13	0.07

Table B.2: US Corrected Item-total Correlations 2025

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.41	0.62	0.60	0.50	0.44
Item 2	0.63	0.53	0.54	0.52	0.54
Item 3	0.70	0.42	0.55	0.50	0.62
Item 4	0.55	0.51	0.68	0.57	0.67
Item 5	0.75	0.62	0.59	0.54	0.47
Item 6	0.53	0.45	0.61	0.40	0.58
Item 7	0.74	0.54	0.50	0.20	0.65
Item 8	0.56	0.60	0.55	0.45	0.56
Item 9	0.74	0.58	0.59	0.45	0.60
Item 10	0.62	0.65	0.55	0.68	0.67
Item 11		0.57	0.58	0.68	0.59
Item 12		0.62	0.65	0.41	
Mean	0.62	0.56	0.58	0.49	0.58
StDev	0.11	0.07	0.05	0.13	0.07

Table B.3: US Corrected Item-total Correlations 2023

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.37	0.62	0.40	0.43	0.50
Item 2	0.38	0.50	0.54	0.53	0.32
Item 3	0.51	0.43	0.43	0.42	0.43
Item 4	0.53	0.58	0.28	0.38	0.37
Item 5	0.59	0.48	0.42	0.51	0.48
Item 6	0.46	0.47	0.34	0.37	0.47
Item 7	0.44	0.37	0.50	0.24	0.51
Item 8	0.61	0.50	0.53	0.42	0.62
Item 9	0.49	0.52	0.60	0.47	0.51
Item 10	0.45	0.34	0.58	0.34	0.49
Item 11		0.53	0.48	0.58	0.30
Item 12		0.48	0.52	0.61	
Mean	0.48	0.48	0.47	0.44	0.45
StDev	0.08	0.08	0.10	0.11	0.09

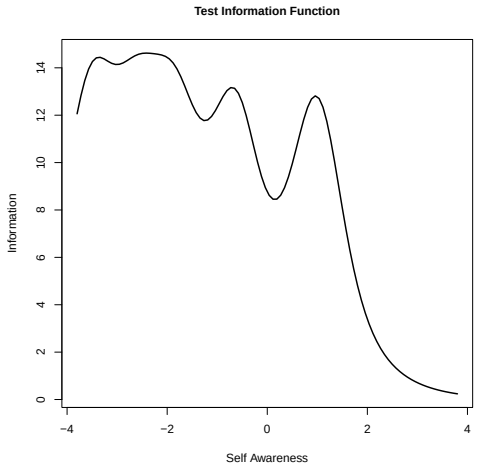


Figure B.1: Self Awareness GRM TIF 2025

B.1.2 US Item Response Theory 2025

Prior to 2025, item response theory (IRT) was only cursorily applied to TTI Success Insights' Emotional Intelligence assessment. As such, there are no results to display in this section. Future versions of this manual will contain historical results as we move forward. For the current results, see Section 11.4.2 above.

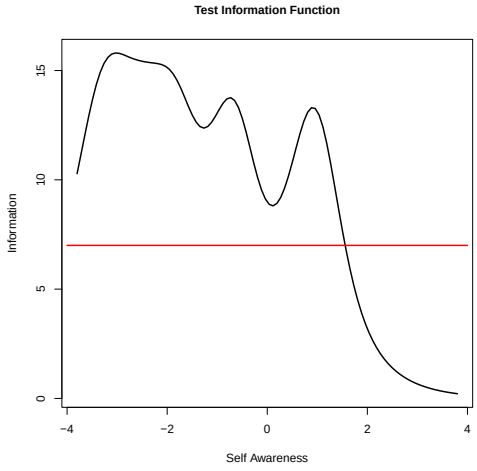


Figure B.2: Self Awareness GRM TIF 2025a

TTI SUCCESS INSIGHTS®

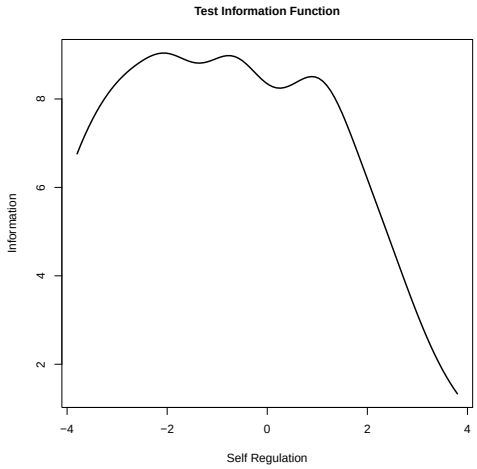


Figure B.3: Self Regulation GRM TIF 2025

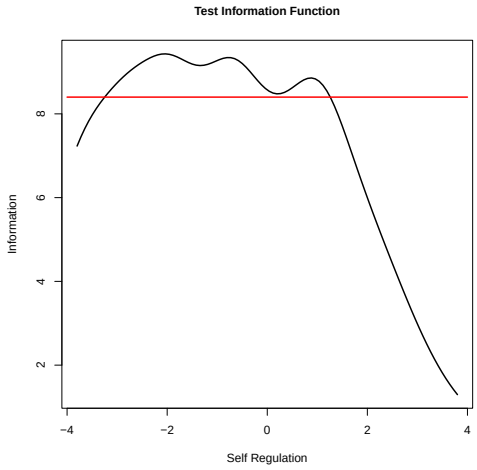


Figure B.4: Self Regulation GRM TIF 2025a

TTI SUCCESS INSIGHTS®

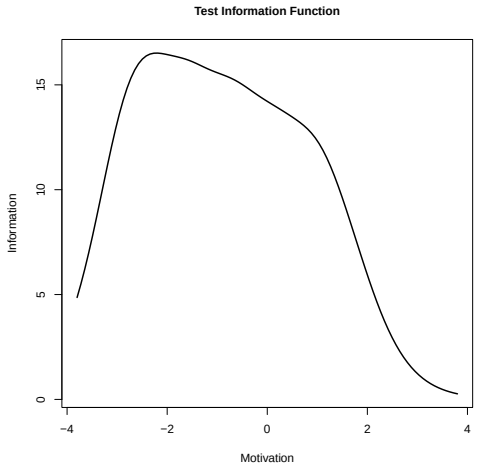


Figure B.5: Motivation GRM TIF 2025

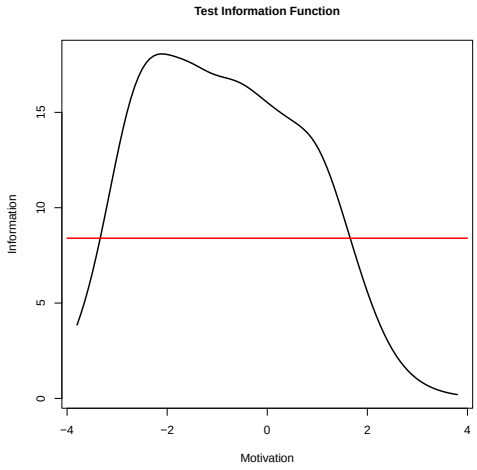


Figure B.6: Motivation GRM TIF 2025a

TTI SUCCESS INSIGHTS®

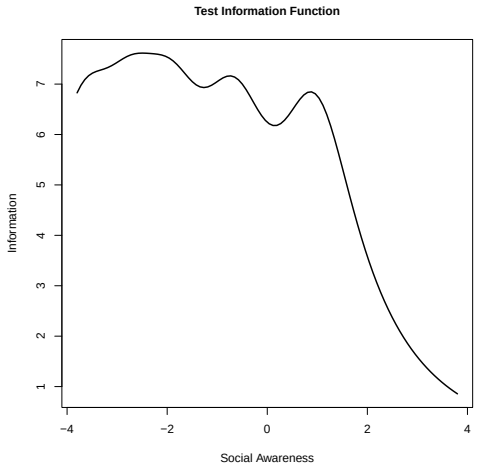


Figure B.7: Social Awareness GRM TIF 2025

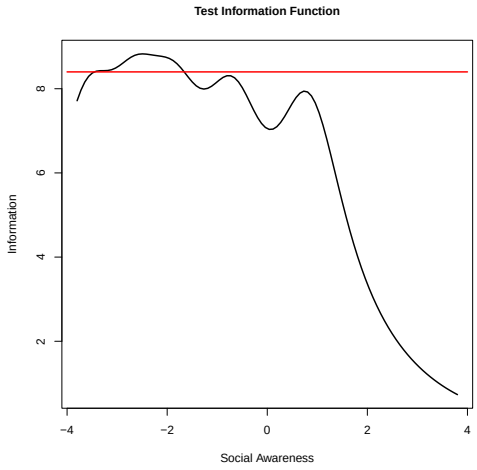


Figure B.8: Social Awareness GRM TIF 2025a

TTI SUCCESS INSIGHTS®

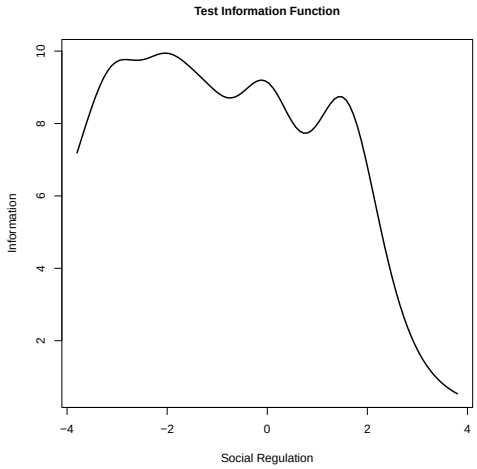


Figure B.9: Social Regulation GRM TIF 2025

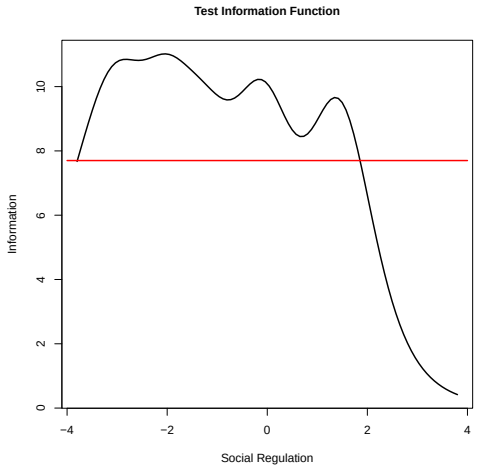


Figure B.10: Social Regulation GRM TIF 2025a

B.1.3 US Factor Analysis 2025

This section presents the results of the factor analysis study for the English U.S. version. The results are presented in Tables B.4 through B.18. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table B.4: US Factor Pattern Matrix
Self Awareness 2025

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.43	
Item 2	0.70		
Item 3	0.64		
Item 4			0.94
Item 5	0.68		
Item 6		0.61	
Item 7	0.72		
Item 8	0.41		
Item 9	0.68	0.40	
Item 10		0.73	
CR	0.89		
VE	0.57		

Table B.5: US Factor Pattern Matrix
Self Awareness 2025

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.43	
Item 2	0.69		
Item 3	0.67		
Item 4			0.44
Item 5	0.61		0.44
Item 6		0.62	
Item 7	0.62		0.43
Item 8			0.63
Item 9	0.55	0.38	0.43
Item 10		0.75	
CR	0.89		
VE	0.55		

TTI SUCCESS INSIGHTS®

Table B.6: US Factor Pattern Matrix
Self Awareness 2023

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.37	
Item 2			0.46
Item 3			0.96
Item 4	0.59		
Item 5	0.58		
Item 6		0.85	
Item 7	0.49		
Item 8	0.71		
Item 9	0.51		
Item 10		0.47	
CR	0.81		
VE	0.46		

Table B.7: US Factor Pattern Matrix
Self Regulation 2025

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.36		0.42	
Item 2				0.78
Item 3			0.58	
Item 4			0.65	
Item 5		0.76		
Item 6			0.48	
Item 7	0.73			
Item 8				0.67
Item 9	0.81			
Item 10		0.83		
Item 11		0.34		
Item 12	0.59		0.36	
CR	0.89			
VE	0.56			

TTI SUCCESS INSIGHTS®

Table B.8: US Factor Pattern Matrix
Self Regulation 2025

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.38		0.38	
Item 2				0.80
Item 3			0.61	
Item 4			0.66	
Item 5		0.76		
Item 6			0.44	
Item 7	0.78			
Item 8				0.68
Item 9	0.80			
Item 10		0.85		
Item 11	0.36			
Item 12	0.61		0.39	
CR	0.89			
VE	0.57			

Table B.9: US Factor Pattern Matrix
Self Regulation 2023

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.47	0.43	
Item 2		0.72		
Item 3	0.55			
Item 4	0.71			
Item 5		0.42		
Item 6	0.35			
Item 7				
Item 8			0.43	
Item 9			0.51	
Item 10				0.49
Item 11			0.41	
Item 12				0.69
CR	0.85			
VE	0.40			

Table B.10: US Factor Pattern Matrix
Motivation 2025

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.80		
Item 2			0.75
Item 3			0.69
Item 4	0.84		
Item 5	0.68		
Item 6	0.63		
Item 7		0.65	
Item 8		0.49	
Item 9		0.71	
Item 10			0.79
Item 11		0.52	0.38
Item 12	0.49	0.43	
CR	0.94		
VE	0.55		

Table B.11: US Factor Pattern Matrix
Motivation 2025

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.82		
Item 2		0.76	
Item 3		0.71	
Item 4	0.85		
Item 5	0.70		
Item 6	0.65		
Item 7			0.65
Item 8			0.51
Item 9			0.72
Item 10		0.70	
Item 11		0.40	0.52
Item 12	0.52		0.44
CR	0.90		
VE	0.57		

Table B.12: US Factor Pattern Matrix
Motivation 2023

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.83			
Item 2	0.76			
Item 3	0.38			
Item 4				
Item 5			0.85	
Item 6				
Item 7		0.69		
Item 8				0.53
Item 9		0.55		
Item 10	0.68			
Item 11				0.44
Item 12			0.37	
CR	0.85			
VE	0.41			

Table B.13: US Factor Pattern Matrix
Social Awareness 2025

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.58	
Item 2		0.81		
Item 3		0.72		
Item 4	0.55			
Item 5	0.45			
Item 6	0.65			
Item 7		0.40		
Item 8			0.85	
Item 9		0.37		
Item 10	0.56			0.47
Item 11	0.51			0.44
Item 12				0.43
CR	0.83			
VE	0.47			

TTI SUCCESS INSIGHTS®

Table B.14: US Factor Pattern Matrix
Social Awareness 2025

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.36		0.49	
Item 2		0.78		
Item 3		0.76		
Item 4	0.59			
Item 5	0.52			
Item 6	0.66			
Item 7		0.38		
Item 8			0.97	
Item 9		0.36		
Item 10	0.59			0.49
Item 11	0.53			0.44
Item 12				0.49
CR	0.86			
VE	0.51			

TTI SUCCESS INSIGHTS®

Table B.15: US Factor Pattern Matrix
Social Awareness 2023

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.37		0.48	
Item 2		0.76		
Item 3	0.42			
Item 4	0.43			
Item 5		0.75		
Item 6	0.56			
Item 7		0.40		
Item 8			0.89	
Item 9				
Item 10				0.85
Item 11	0.62			
Item 12	0.57			
CR	0.82			
VE	0.46			

Table B.16: US Factor Pattern Matrix
Social Regulation 2025

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.86
Item 2		0.62		
Item 3			0.81	
Item 4	0.53		0.38	
Item 5				0.82
Item 6	0.77			
Item 7	0.77			
Item 8		0.85		
Item 9	0.52			
Item 10	0.37		0.77	
Item 11		0.56		
CR	0.87			
VE	0.64			

Table B.17: US Factor Pattern Matrix
Social Regulation 2025

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.78
Item 2		0.66		
Item 3			0.79	
Item 4	0.62			
Item 5				0.89
Item 6	0.77			
Item 7	0.76			
Item 8		0.85		
Item 9	0.60			
Item 10	0.43		0.75	
Item 11		0.60		
CR	0.89			
VE	0.67			

Table B.18: US Factor Pattern Matrix
Social Regulation 2023

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.40		
Item 2				
Item 3			0.50	
Item 4		0.47		
Item 5				0.93
Item 6		0.46		
Item 7	0.66			
Item 8		0.39	0.59	
Item 9	0.63			
Item 10	0.54			
Item 11			0.40	
CR	0.82			
VE	0.42			

C EQ Analysis for Australia

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Australian version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 3380 respondents. Table C.1 presents the general statistics on means and standard deviations for the Australian population, as well as for a Male/Female breakdown.

Table C.1: Australian General Population Statistics
N = 3380

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	74.98	10.60	73.17	10.55	76.79	10.34
Self Regulation	67.91	11.90	69.09	11.68	66.72	12.00
Motivation	74.19	11.28	74.24	11.12	74.14	11.43
Social Awareness	71.73	11.57	69.05	11.74	74.41	10.74
Social Regulation	69.21	11.52	68.39	11.56	70.04	11.43

C.1 Australian Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the English-Australian version. The results are presented in Table C.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table C.2: Average Australian Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.33	0.25	0.28	0.27	0.27
Item 2	0.32	0.26	0.27	0.25	0.29
Item 3	0.30	0.27	0.27	0.26	0.28
Item 4	0.30	0.26	0.29	0.27	0.28
Item 5	0.29	0.26	0.28	0.25	0.27
Item 6	0.32	0.26	0.28	0.27	0.27
Item 7	0.31	0.27	0.27	0.28	0.27
Item 8	0.29	0.26	0.27	0.27	0.26
Item 9	0.31	0.27	0.26	0.26	0.27
Item 10	0.32	0.28	0.26	0.27	0.27
Item 11		0.26	0.27	0.25	0.30
Item 12		0.28	0.27	0.25	
Mean	0.31	0.26	0.27	0.26	0.28
StDev	0.01	0.01	0.01	0.01	0.01

TTI SUCCESS INSIGHTS®

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Australian version based on inter-item correlation.

C.2 Australian Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the English-Australian version. The results are presented in Table C.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table C.3: Australian Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.37	0.63	0.40	0.41	0.61
Item 2	0.42	0.50	0.53	0.57	0.39
Item 3	0.55	0.40	0.45	0.44	0.48
Item 4	0.53	0.56	0.28	0.42	0.45
Item 5	0.62	0.47	0.43	0.56	0.54
Item 6	0.46	0.49	0.38	0.36	0.54
Item 7	0.48	0.37	0.50	0.28	0.57
Item 8	0.61	0.49	0.50	0.42	0.68
Item 9	0.48	0.45	0.57	0.48	0.60
Item 10	0.44	0.32	0.57	0.40	0.58
Item 11		0.49	0.48	0.55	0.33
Item 12		0.35	0.51	0.60	
Mean	0.50	0.46	0.47	0.46	0.52
StDev	0.08	0.10	0.08	0.10	0.10

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table C.3, we have highlighted potential problem areas in red. These two items are identified for monitoring and possible replacement at some point in the future. There is no alarm to be sounded at this time. The mean item-to-total correlation scores are are in an acceptable range and the standard deviations show consistent scoring.

C.3 Australian Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the English-Australian version. The results are presented in Table C.5. For convenience, Table 6.1 is reproduced as Table C.4.

Table C.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table C.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table C.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table C.5: Average Australian α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.81	0.81	0.80	0.81	0.79
Lower CI	0.80	0.80	0.79	0.80	0.78
Upper CI	0.82	0.82	0.81	0.82	0.80

As is presented in Table C.5, all α values are in the Very Good ranges with the lone exception being the dip to 0.79 for the Social Regulation scale, implying a Respectable internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Australian version of EQ has solid evidence of internal consistency reliability.

Table C.6: α-if-item-deleted for Australia

Est./Scale	SA	SR	MO	SoA	SoR
α	0.82	0.81	0.82	0.81	0.81
Item 1	0.82	0.78	0.81	0.80	0.78
Item 2	0.81	0.80	0.80	0.78	0.80
Item 3	0.80	0.80	0.81	0.80	0.80
Item 4	0.80	0.79	0.82	0.80	0.80
Item 5	0.79	0.80	0.81	0.79	0.79
Item 6	0.81	0.80	0.81	0.80	0.79
Item 7	0.80	0.81	0.80	0.81	0.79
Item 8	0.79	0.80	0.80	0.80	0.78
Item 9	0.80	0.80	0.80	0.79	0.79
Item 10	0.81	0.81	0.80	0.80	0.79
Item 11		0.80	0.80	0.79	0.81
Item 12		0.81	0.80	0.78	

C.4 Australian Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

C.5 Australian Factor Analysis

This section presents the results of the factor analysis study for the Australian version. The results are presented in Tables C.7 through C.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table C.7: Australian Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1			0.42
Item 2		0.98	
Item 3		0.51	
Item 4	0.58		
Item 5	0.60		
Item 6			0.80
Item 7	0.53		
Item 8	0.73		
Item 9	0.49		
Item 10			0.46
CR	0.82		
VE	0.48		

Some comments on how to interpret the information presented in Tables C.7 through C.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table C.8: Australian Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.46	0.42	
Item 2			0.73	
Item 3	0.53			
Item 4	0.74			
Item 5			0.43	
Item 6	0.40			
Item 7		0.35		
Item 8		0.40		
Item 9		0.46		
Item 10				0.57
Item 11		0.51		
Item 12				0.53
CR	0.84			
VE	0.48			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table C.9: Australian Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2	0.79			
Item 3	0.41			
Item 4				
Item 5			0.83	
Item 6				
Item 7				0.66
Item 8		0.52		
Item 9		0.49		0.50
Item 10	0.69			
Item 11		0.45		
Item 12			0.40	
CR	0.83			
VE	0.41			

TTI SUCCESS INSIGHTS®

Table C.10: Australian Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.97	
Item 2	0.80			
Item 3		0.42		
Item 4		0.49		
Item 5	0.74			
Item 6		0.56		
Item 7	0.40			
Item 8		0.37	0.43	
Item 9	0.42	0.30		
Item 10	0.35			0.71
Item 11		0.55		
Item 12	0.34	0.53		
CR	0.83			
VE	0.46			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table C.11: Australian Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.44		
Item 2		0.35		
Item 3				0.47
Item 4		0.48		
Item 5			0.93	
Item 6		0.47		
Item 7	0.62			
Item 8		0.45		0.48
Item 9	0.70			
Item 10	0.57			
Item 11				0.42
CR	0.83			
VE	0.43			

The main takeaway for the Australian EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation



and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

C.6 Australian EQ Model Equating Functions

Table C.12: Australian-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.98	3.32	1.00	2.54	0.98	3.35
Self Regulation	1.02	1.62	1.02	2.05	1.02	1.36
Motivation	0.99	3.60	0.99	3.89	0.98	3.56
Social Awareness	1.00	0.58	1.01	-0.01	0.99	1.41
Social Regulation	1.00	1.78	1.02	1.33	0.99	2.07

Table C.13: Australian-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.94	3.14	0.93	3.20	0.96	1.67
Self Regulation	0.94	1.61	0.93	2.21	0.94	1.83
Motivation	0.96	0.52	0.92	3.26	0.98	-0.43
Social Awareness	1.00	-0.73	0.98	0.60	0.94	3.58
Social Regulation	0.95	1.93	0.95	1.31	0.95	2.25

C.7 Australian EQ Model Summary

The Australian version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Australian model.

D EQ Analysis for Brazil

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Brazilian version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 4530 respondents. Table D.1 presents the general statistics on means and standard deviations for the Brazilian population, as well as for a Male/Female breakdown.

Table D.1: Brazilian General Population Statistics
N = 4530

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	68.30	9.67	66.93	9.36	69.67	9.79
Self Regulation	67.52	12.97	69.77	12.18	65.28	13.34
Motivation	74.58	11.71	76.29	11.17	72.86	11.99
Social Awareness	66.75	11.33	64.69	11.42	68.81	10.85
Social Regulation	71.71	11.74	72.08	11.79	71.33	11.68

D.1 Brazilian Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Brazilian version. The results are presented in Table D.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table D.2: Average Brazilian Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.14	0.24	0.26	0.19	0.23
Item 2	0.15	0.25	0.25	0.17	0.24
Item 3	0.14	0.26	0.26	0.17	0.23
Item 4	0.19	0.27	0.28	0.18	0.24
Item 5	0.18	0.26	0.26	0.17	0.23
Item 6	0.13	0.24	0.27	0.17	0.23
Item 7	0.15	0.28	0.26	0.19	0.22
Item 8	0.18	0.26	0.25	0.18	0.22
Item 9	0.14	0.26	0.24	0.17	0.22
Item 10	0.14	0.27	0.25	0.19	0.23
Item 11		0.25	0.27	0.17	0.24
Item 12		0.27	0.25	0.16	
Mean	0.15	0.26	0.26	0.18	0.23
StDev	0.02	0.01	0.01	0.01	0.01

The results are quite solid for three of the five scales. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Self Awareness and Social Awareness values indicate a need to identify each construct, in its entirety, for monitoring and probable replacement. It is noted that an update to the development language questionnaire is forthcoming and may alleviate some of the issues identified in this analysis.

D.2 Brazilian Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Brazilian version. The results are presented in Table D.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table D.3: Brazilian Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.41	0.65	0.47	0.25	0.38
Item 2	0.37	0.56	0.50	0.44	0.34
Item 3	0.44	0.47	0.41	0.37	0.45
Item 4	0.05	0.34	0.30	0.30	0.30
Item 5	0.13	0.50	0.41	0.38	0.44
Item 6	0.46	0.63	0.30	0.40	0.38
Item 7	0.30	0.26	0.41	0.22	0.47
Item 8	0.06	0.42	0.49	0.32	0.53
Item 9	0.40	0.42	0.63	0.41	0.48
Item 10	0.40	0.36	0.55	0.22	0.43
Item 11		0.54	0.34	0.43	0.35
Item 12		0.34	0.55	0.46	
Mean	0.30	0.46	0.45	0.35	0.41
StDev	0.16	0.12	0.11	0.09	0.07

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table D.3, we have highlighted potential problem areas in red. The Self Awareness and Social Awareness scales show behavior consistent with that identified in the inter-item correlation analysis. These scales are marked for continued monitoring.

D.3 Brazilian Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Brazilian version. The results are presented in Table D.5. For convenience, Table 6.1 is reproduced as Table D.4.

Table D.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table D.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table D.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table D.5: Average Brazilian α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.61	0.81	0.79	0.72	0.76
Lower CI	0.59	0.80	0.78	0.71	0.75
Upper CI	0.63	0.82	0.80	0.73	0.77

As is presented in Table D.5, all α values are in the Very Good to Respectable ranges with the lone exception being the dip to 0.61 for the Self Awareness scale, implying an Undesirable internal consistency. This is consistent with earlier results.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are items identified as having an increase in the value of α at this time. We conclude that the Australian version of EQ has some evidence of internal consistency reliability on four of the five scales.

Table D.6: α-if-item-deleted for Australia

Est./Scale	SA	SR	MO	SoA	SoR
α	0.61	0.81	0.79	0.72	0.76
Item 1	0.56	0.78	0.77	0.71	0.74
Item 2	0.56	0.79	0.77	0.68	0.74
Item 3	0.54	0.80	0.77	0.69	0.73
Item 4	0.65	0.81	0.79	0.70	0.75
Item 5	0.62	0.79	0.78	0.69	0.73
Item 6	0.55	0.78	0.78	0.68	0.74
Item 7	0.58	0.81	0.78	0.71	0.73
Item 8	0.64	0.80	0.77	0.70	0.73
Item 9	0.56	0.80	0.75	0.68	0.73
Item 10	0.56	0.80	0.76	0.71	0.73
Item 11		0.79	0.79	0.68	0.75
Item 12		0.80	0.76	0.68	

D.4 Brazilian Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

D.5 Brazilian Factor Analysis

This section presents the results of the factor analysis study for the Australian version. The results are presented in Tables D.7 through D.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table D.7: Brazilian Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.54		
Item 2		0.74	
Item 3		0.73	
Item 4			0.65
Item 5			0.55
Item 6	0.79		
Item 7			
Item 8			
Item 9	0.39		
Item 10	0.60		
CR	0.61		
VE	0.37		

Some comments on how to interpret the information presented in Tables D.7 through D.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table D.8: Brazilian Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.82			
Item 2	0.37	0.54		
Item 3			0.56	
Item 4			0.48	
Item 5		0.68		
Item 6		0.39	0.45	
Item 7				0.42
Item 8				0.34
Item 9				
Item 10				0.47
Item 11		0.47		
Item 12				0.54
CR	0.84			
VE	0.40			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table D.9: Brazilian Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.43			
Item 2		0.66		
Item 3				
Item 4				
Item 5			0.83	
Item 6				
Item 7	0.54			
Item 8				0.70
Item 9	0.58			0.39
Item 10		0.68		
Item 11				
Item 12	0.46		0.40	
CR	0.84			
VE	0.39			

TTI SUCCESS INSIGHTS®

Table D.10: Brazilian Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.55	
Item 2	0.64			
Item 3		0.41		
Item 4		0.52		
Item 5	0.68			
Item 6		0.53		
Item 7	0.45			
Item 8			0.94	
Item 9	0.42			
Item 10				
Item 11		0.58		
Item 12	0.40	0.37		
CR	0.75			
VE	0.37			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table D.11: Brazilian Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2				0.46
Item 3			0.60	
Item 4				
Item 5		0.45		
Item 6		0.60		
Item 7	0.53			
Item 8		0.37	0.44	
Item 9	0.69			
Item 10	0.57			
Item 11			0.43	
CR	0.79			
VE	0.35			

The main takeaway for the Brazilian EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good for three of the five scales. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this



situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all EQ models at TTI Success Insights.

D.6 Brazilian EQ Model Equating Functions

Table D.12: Brazilian-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.08	3.47	1.12	0.30	1.03	6.47
Self Regulation	0.94	7.71	0.97	4.30	0.92	9.56
Motivation	0.95	5.93	0.99	2.16	0.93	8.15
Social Awareness	1.02	4.14	1.04	2.57	0.98	7.67
Social Regulation	0.99	0.62	1.00	-1.00	0.97	2.30

Table D.13: Brazilian-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.11	1.12	1.16	-2.17	1.08	3.26
Self Regulation	0.95	6.89	0.99	3.05	0.93	8.73
Motivation	0.99	3.39	1.05	-2.85	0.96	5.93
Social Awareness	1.06	-2.67	1.08	-4.00	1.08	-3.55
Social Regulation	1.01	-1.18	0.99	-0.54	1.01	-0.84

D.7 Brazilian EQ Model Summary

The Brazilian version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Brazilian model.

E EQ Analysis for Chinese

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Chinese version of the TTI Success Insights EQ assessment. The study is conducted on 1434 respondents. Due to the relatively small sample size used for this study, some results may be inconsistent. able E.1 presents the general statistics on means and standard deviations for the Chinese population, as well as for a Male/Female breakdown.

Table E.1: Chinese General Population Statistics
N = 1434

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	74.16	10.81	73.57	11.10	74.75	10.49
Self Regulation	61.66	11.20	63.60	11.03	59.72	11.04
Motivation	69.65	11.30	71.20	11.19	68.10	11.20
Social Awareness	68.12	9.77	68.23	9.91	68.00	9.63
Social Regulation	63.16	11.19	63.84	11.14	62.49	11.20

E.1 Chinese Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Chinese version. The results are presented in Table E.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table E.2: Average Chinese Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.31	0.27	0.34	0.17	0.23
Item 2	0.33	0.28	0.31	0.15	0.22
Item 3	0.31	0.31	0.32	0.17	0.22
Item 4	0.29	0.30	0.34	0.19	0.22
Item 5	0.27	0.29	0.33	0.15	0.24
Item 6	0.29	0.27	0.32	0.18	0.22
Item 7	0.28	0.28	0.32	0.16	0.22
Item 8	0.27	0.27	0.31	0.17	0.21
Item 9	0.29	0.27	0.30	0.19	0.22
Item 10	0.30	0.30	0.31	0.17	0.21
Item 11		0.27	0.34	0.16	0.26
Item 12		0.29	0.32	0.15	
Average	0.30	0.28	0.32	0.17	0.22
StDev	0.02	0.01	0.01	0.01	0.02

The results are quite solid overall, with the exception being the Social Awareness construct. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Chinese version based on inter-item correlation in four of the five constructs. The entire Social Awareness construct is marked for monitoring and likely replacement of many during the adaptation process related to the upcoming update to the development language questionnaire.

E.2 Chinese Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Chinese version. The results are presented in Table E.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table E.3: Chinese Corrected Item-total Correlations
TTI SI

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.37	0.60	0.34	0.31	0.33
Item 2	0.26	0.51	0.60	0.46	0.43
Item 3	0.40	0.28	0.55	0.30	0.48
Item 4	0.52	0.33	0.38	0.15	0.48
Item 5	0.62	0.45	0.45	0.45	0.28
Item 6	0.48	0.56	0.49	0.18	0.42
Item 7	0.57	0.48	0.55	0.36	0.49
Item 8	0.63	0.59	0.65	0.35	0.50
Item 9	0.48	0.59	0.69	0.09	0.47
Item 10	0.45	0.28	0.63	0.34	0.54
Item 11		0.62	0.34	0.45	0.41
Item 12		0.43	0.54	0.51	
Average	0.48	0.47	0.53	0.33	0.41
StDev	0.12	0.12	0.11	0.14	0.13

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table E.3, we have highlighted potential problem areas in red. Consistent with the inter-item correlation results, we see some significant issues in the Social Awareness construct. In particular,



we see some fairly low correlations on this scale, implying those items are not consistent with the intent of the construct.

E.3 Chinese Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Chinese version. The results are presented in Table E.5. For convenience, Table 6.1 is reproduced as Table E.4.

Table E.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table E.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table E.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table E.5: Average Chinese α Coefficient Data TTI SI

Est./Scale	SA	SR	MO	SoA	SoR
α	0.78	0.81	0.84	0.66	0.75
Lower CI	0.77	0.80	0.83	0.63	0.74
Upper CI	0.79	0.83	0.85	0.69	0.77

As is presented in Table E.5, α values for two of the five constructs are in the Very Good range. Self Awareness and Social Regulation drop into the Respectable range. Social Awareness falls to the Undesirable range, just above the Unacceptable range, consistent with results already seen above.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Chinese version of EQ has solid evidence of internal consistency reliability for all constructs outside Social Awareness.

Table E.6: α-if-item-deleted for Chinese

Est./Scale	SA	SR	MO	SoA	SoR
α	0.81	0.83	0.85	0.71	0.76
Item 1	0.78	0.79	0.84	0.64	0.75
Item 2	0.80	0.80	0.82	0.62	0.73
Item 3	0.78	0.82	0.82	0.64	0.73
Item 4	0.76	0.81	0.84	0.68	0.73
Item 5	0.75	0.80	0.83	0.62	0.75
Item 6	0.77	0.79	0.83	0.66	0.74
Item 7	0.76	0.80	0.82	0.63	0.73
Item 8	0.75	0.79	0.82	0.63	0.72
Item 9	0.77	0.79	0.81	0.69	0.73
Item 10	0.77	0.82	0.82	0.64	0.72
Item 11		0.79	0.84	0.62	0.78
Item 12		0.80	0.82	0.62	

E.4 Chinese Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

E.5 Chinese Factor Analysis

This section presents the results of the factor analysis study for the Chinese version. The results are presented in Tables E.7 through E.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table E.7: Chinese Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.33		
Item 2		0.37	
Item 3		0.98	
Item 4	0.58		
Item 5	0.60		
Item 6			0.79
Item 7	0.68		
Item 8	0.73		
Item 9	0.56		
Item 10	0.42		
CR	0.81		
VE	0.47		

TTI SUCCESS INSIGHTS®

Some comments on how to interpret the information presented in Tables E.7 through E.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor pattern matrix, i.e., there is little or no cross-loading present.

Table E.8: Chinese Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.70			
Item 2	0.45			
Item 3				0.55
Item 4	-0.47			
Item 5			0.93	
Item 6		0.46		0.43
Item 7	0.40	0.41		
Item 8		0.68		
Item 9		0.58		
Item 10				
Item 11	0.60	0.38		
Item 12		0.35		
CR	0.82			
VE	0.45			

Cross-loading is a reference to each item having a nontrivial loading on more than one factor.

Table E.9: Chinese Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2		0.91		
Item 3	0.47			
Item 4				
Item 5			0.97	
Item 6	0.51			
Item 7	0.66			
Item 8	0.71			
Item 9	0.77			
Item 10		0.49		0.80
Item 11				
Item 12	0.56			
CR	0.85			
VE	0.52			

TTI SUCCESS INSIGHTS®

Table E.10: Chinese Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.76		
Item 2	0.44			
Item 3				
Item 4				
Item 5	0.67			
Item 6			0.43	
Item 7	0.57			
Item 8		0.69		
Item 9				
Item 10	0.57			
Item 11	0.50		0.46	
Item 12	0.38			0.64
CR	0.74			
VE	0.37			

Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it. Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions. The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table E.11: Chinese Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2		0.37		
Item 3				0.58
Item 4		0.66		
Item 5			0.98	
Item 6		0.43		
Item 7	0.52			
Item 8				0.42
Item 9	0.62			
Item 10	0.63			
Item 11				
CR	0.79			
VE	0.41			

TTI SUCCESS INSIGHTS®

The main takeaway for the Chinese EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

E.6 Chinese EQ Model Equating Functions

Table E.12: Chinese-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.05	-0.85	1.03	-0.79	1.02	2.19
Self Regulation	1.11	2.17	1.12	1.09	1.14	1.23
Motivation	1.04	4.02	1.06	1.81	1.05	4.26
Social Awareness	1.21	-10.24	1.25	-15.31	1.15	-3.08
Social Regulation	1.08	2.71	1.09	0.88	1.05	5.67

Table E.13: Chinese-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.96	5.54	0.95	5.77	0.96	6.38
Self Regulation	1.08	4.06	1.08	3.83	1.11	3.21
Motivation	0.99	8.20	0.99	7.34	1.00	8.11
Social Awareness	1.19	-8.39	1.20	-11.93	1.11	-0.00
Social Regulation	1.00	1.78	1.02	1.33	0.99	2.07

E.7 Chinese EQ Model Summary

The Chinese version of the EQ model shows reasonable evidence of internal consistency reliability and internal structure validity in four of the five constructs. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Chinese model.

F EQ Analysis for Dutch

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Dutch version of the TTI Success Insights EQ assessment. The study is conducted on 1647 respondents. Due to a relatively small subset of data, some results may not be as reliable as desired at this time. Table F.1 presents the general statistics on means and standard deviations for the Australian population, as well as for a Male/Female breakdown.

Table F.1: Dutch General Population Statistics
N = 1647

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	74.81	10.56	73.64	10.74	75.98	10.25
Self Regulation	64.61	11.45	66.36	11.39	62.85	11.24
Motivation	72.89	11.52	73.92	11.56	71.86	11.40
Social Awareness	70.60	12.01	67.19	12.44	74.02	10.51
Social Regulation	69.72	11.71	68.36	12.41	71.09	10.80

F.1 Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Dutch version. The results are presented in Table F.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table F.2: Average Dutch Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.27	0.19	0.25	0.30	0.25
Item 2	0.27	0.20	0.25	0.28	0.28
Item 3	0.25	0.23	0.26	0.30	0.26
Item 4	0.28	0.21	0.27	0.31	0.26
Item 5	0.28	0.20	0.24	0.29	0.27
Item 6	0.26	0.21	0.27	0.32	0.26
Item 7	0.27	0.22	0.24	0.32	0.26
Item 8	0.27	0.22	0.24	0.31	0.25
Item 9	0.26	0.21	0.24	0.30	0.26
Item 10	0.27	0.23	0.24	0.31	0.25
Item 11		0.20	0.25	0.27	0.29
Item 12		0.22	0.26	0.29	
Average	0.27	0.21	0.25	0.30	0.26
StDev	0.01	0.01	0.01	0.02	0.01

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Australian version based on inter-item correlation. We do highlight one item in orange for falling below the 0.20 value. This item should be monitored for the time being.

F.2 Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Dutch version. The results are presented in Table F.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure

Table F.3: Dutch Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.40	0.58	0.43	0.48	0.59
Item 2	0.46	0.53	0.48	0.69	0.32
Item 3	0.57	0.27	0.38	0.50	0.44
Item 4	0.36	0.41	0.29	0.41	0.48
Item 5	0.41	0.48	0.50	0.60	0.35
Item 6	0.49	0.44	0.29	0.35	0.49
Item 7	0.45	0.35	0.50	0.32	0.46
Item 8	0.47	0.30	0.54	0.39	0.58
Item 9	0.47	0.38	0.49	0.52	0.46
Item 10	0.45	0.21	0.58	0.36	0.53
Item 11		0.54	0.47	0.72	0.26
Item 12		0.30	0.38	0.61	
Average	0.45	0.37	0.44	0.48	0.44
StDev	0.06	0.10	0.10	0.13	0.10

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table F.3, we have highlighted potential problem areas in red. There are several items on the Self Regulation construct, in particular, that may need some attention. There are also two items on the Motivation construct and one on the Social Regulation construct that are questionable. As noted in other areas of this report, there is a scheduled update to the U.S. English based questionnaire that will provide many changes. Eventually, the questionnaire will be adapted to Dutch and we shall revisit these issues at that time.

F.3 Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Dutch version. The results are presented in Table F.5. For convenience, Table 6.1 is reproduced as Table F.4.

Table F.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table F.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table F.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table F.5: Average Dutch α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.78	0.74	0.79	0.83	0.79
Lower CI	0.78	0.76	0.78	0.82	0.77
Upper CI	0.80	0.77	0.81	0.84	0.80

As is presented in Table F.5, all α values are in the Respectable range. The exception is the Social Awareness construct where the α values rise into the Very Good range. All these numbers are solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Dutch version of EQ has solid evidence of internal consistency reliability.

Table F.6: α-if-item-deleted for Dutch

Est./Scale	SA	SR	MO	SoA	SoR
α	0.79	0.76	0.80	0.84	0.79
Item 1	0.77	0.73	0.79	0.83	0.77
Item 2	0.77	0.73	0.78	0.81	0.79
Item 3	0.75	0.76	0.79	0.82	0.78
Item 4	0.78	0.75	0.80	0.83	0.78
Item 5	0.77	0.75	0.78	0.81	0.79
Item 6	0.76	0.74	0.80	0.84	0.77
Item 7	0.77	0.75	0.78	0.84	0.78
Item 8	0.76	0.76	0.78	0.83	0.77
Item 9	0.76	0.75	0.78	0.82	0.78
Item 10	0.77	0.77	0.77	0.83	0.77
Item 11		0.73	0.78	0.81	0.80
Item 12		0.76	0.79	0.82	

F.4 Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

F.5 Factor Analysis

This section presents the results of the factor analysis study for the Dutch version. The results are presented in Tables F.7 through F.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table F.7: Dutch Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.32		
Item 2	0.59		
Item 3	0.55		
Item 4			0.54
Item 5			0.67
Item 6		0.91	
Item 7	0.43		
Item 8	0.60		
Item 9	0.49		
Item 10	0.40	0.37	
CR	0.79		
VE	0.40		

Some comments on how to interpret the information presented in Tables F.7 through F.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table F.8: Dutch Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.64			
Item 2	0.78			
Item 3		0.45		
Item 4		0.68		
Item 5	0.61			
Item 6				
Item 7			0.37	
Item 8				0.36
Item 9				0.51
Item 10				
Item 11	0.52			
Item 12			0.45	
CR	0.76			
VE	0.35			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table F.9: Dutch Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.38			
Item 2		0.65		0.39
Item 3				
Item 4				
Item 5			0.85	
Item 6				
Item 7	0.64			
Item 8				0.48
Item 9	0.53			
Item 10		0.73		
Item 11				0.61
Item 12			0.46	
CR	0.80			
VE	0.40			

Table F.10: Dutch Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.75		
Item 2	0.69		0.42	
Item 3			0.41	
Item 4			0.49	
Item 5	0.74			
Item 6			0.46	
Item 7	0.51			
Item 8		0.76		
Item 9	0.47			
Item 10				0.94
Item 11	0.77		0.35	
Item 12	0.43		0.39	
CR	0.74			
VE	0.37			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table F.11: Dutch Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.38	0.45		
Item 2				
Item 3		0.66		
Item 4	0.62			
Item 5				0.63
Item 6	0.59			
Item 7			0.58	
Item 8	0.36	0.57		
Item 9			0.76	
Item 10	0.41		0.35	
Item 11				
CR	0.79			
VE	0.41			

The main takeaway for the Dutch EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation



and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all EQ models at TTI Success Insights.

F.6 Dutch EQ Model Equating Functions

Table F.12: Dutch-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.98	3.21	0.98	3.37	0.98	3.49
Self Regulation	1.06	2.40	1.04	3.11	1.09	0.98
Motivation	0.97	6.38	0.96	6.99	0.98	5.61
Social Awareness	0.97	4.31	0.95	5.69	1.01	0.26
Social Regulation	0.99	2.40	0.95	6.13	1.05	-3.08

Table F.13: Dutch-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.06	-2.71	1.06	-2.49	1.04	-0.51
Self Regulation	1.07	1.26	1.08	0.24	1.11	-0.20
Motivation	1.03	1.69	1.03	1.63	1.04	1.39
Social Awareness	0.99	2.81	0.99	3.18	1.07	-4.46
Social Regulation	1.03	-1.03	0.98	3.90	1.09	-6.23

F.7 Dutch EQ Model Summary

The Dutch version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Dutch model.

G EQ Analysis for French

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the French (France) version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 4146 respondents. Table G.1 presents the general statistics on means and standard deviations for the Australian population, as well as for a Male/Female breakdown.

Table G.1: French General Population Statistics
N = 4146

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	67.02	9.38	65.59	9.11	68.44	9.43
Self Regulation	65.03	11.66	66.92	11.08	63.13	11.92
Motivation	74.91	10.10	75.09	10.07	74.73	10.14
Social Awareness	66.75	11.33	64.69	11.42	68.81	10.85
Social Regulation	67.49	10.87	67.75	11.09	67.23	10.64

G.1 French Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the French version. The results are presented in Table G.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table G.2: Average French Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.14	0.20	0.21	0.20	0.21
Item 2	0.15	0.22	0.20	0.19	0.21
Item 3	0.14	0.22	0.21	0.21	0.20
Item 4	0.18	0.20	0.22	0.22	0.21
Item 5	0.17	0.21	0.21	0.20	0.20
Item 6	0.14	0.21	0.24	0.20	0.20
Item 7	0.16	0.23	0.20	0.22	0.20
Item 8	0.18	0.22	0.19	0.20	0.20
Item 9	0.15	0.21	0.20	0.20	0.20
Item 10	0.14	0.23	0.19	0.22	0.20
Item 11		0.21	0.20	0.19	0.22
Item 12		0.23	0.20	0.19	
Mean	0.15	0.22	0.20	0.20	0.20
StDev	0.02	0.01	0.01	0.01	0.01

The results are questionable overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since four of the five scales are at or below 0.20 for an average inter-item correlation score, the entire French EQ model needs to be addressed.

G.2 French Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the French version. The results are presented in Table G.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table G.3: French Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.38	0.61	0.36	0.43	0.36
Item 2	0.34	0.37	0.47	0.47	0.29
Item 3	0.46	0.40	0.31	0.34	0.39
Item 4	0.11	0.55	0.27	0.26	0.29
Item 5	0.20	0.43	0.34	0.43	0.43
Item 6	0.44	0.52	0.08	0.41	0.39
Item 7	0.27	0.25	0.41	0.20	0.46
Item 8	0.07	0.35	0.51	0.37	0.48
Item 9	0.35	0.42	0.45	0.45	0.44
Item 10	0.39	0.25	0.51	0.25	0.45
Item 11		0.51	0.46	0.55	0.23
Item 12		0.24	0.41	0.50	
Mean	0.30	0.41	0.38	0.39	0.38
StDev	0.13	0.12	0.12	0.11	0.08

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table G.3, we have highlighted potential problem areas in red. The Self Awareness scale shows results highly consistent with those of the previous section. The relatively low average item-to-total correlations are indicative of a model in need of attention.

G.3 French Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the French version. The results are presented in Table G.5. For convenience, Table 6.1 is reproduced as Table G.4.

Table G.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table G.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table G.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table G.5: Average Frencha α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.61	0.77	0.73	0.75	0.73
Lower CI	0.59	0.76	0.71	0.74	0.72
Upper CI	0.63	0.78	0.74	0.76	0.74

As is presented in Table G.5, α values are in the Respectable range with the lone exception being the dip to 0.61 for the Self Awareness scale, implying an Unacceptable internal consistency. This result for Self Awareness is consistent with earlier findings.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the French version of EQ has reasonable evidence of internal consistency reliability in four of the five scales.

Table G.6: α-if-item-deleted for France

Est./Scale	SA	SR	MO	SoA	SoR
α	0.61	0.77	0.73	0.75	0.81
Item 1	0.59	0.73	0.71	0.73	0.71
Item 2	0.57	0.76	0.70	0.72	0.72
Item 3	0.54	0.76	0.71	0.74	0.71
Item 4	0.63	0.74	0.74	0.75	0.72
Item 5	0.61	0.75	0.71	0.73	0.70
Item 6	0.56	0.75	0.75	0.73	0.71
Item 7	0.59	0.77	0.71	0.75	0.70
Item 8	0.65	0.76	0.69	0.74	0.69
Item 9	0.57	0.75	0.70	0.72	0.70
Item 10	0.57	0.77	0.69	0.75	0.70
Item 11		0.74	0.70	0.71	0.73
Item 12		0.77	0.70	0.72	

G.4 French Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

G.5 French Factor Analysis

This section presents the results of the factor analysis study for the French version. The results are presented in Tables G.7 through G.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table G.7: French Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.46		
Item 2		0.62	
Item 3		0.83	
Item 4			0.75
Item 5			0.46
Item 6	0.79		
Item 7			
Item 8			
Item 9	0.36		
Item 10	0.59		
CR	0.62		
VE	0.36		

Some comments on how to interpret the information presented in Tables G.7 through G.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table G.8: French Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.66			
Item 2			0.41	
Item 3	0.43			
Item 4		0.76		
Item 5			0.54	
Item 6	0.39			
Item 7				0.50
Item 8				
Item 9				
Item 10				
Item 11	0.52			
Item 12				
CR	0.80			
VE	0.33			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table G.9: French Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2		0.64		
Item 3		0.45		
Item 4			0.98	
Item 5				0.64
Item 6				
Item 7	0.49			
Item 8	0.57			
Item 9	0.58			
Item 10		0.53		
Item 11	0.51			
Item 12	0.35		0.38	
CR	0.79			
VE	0.38			

TTI SUCCESS INSIGHTS®

Table G.10: French Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.41	0.52	
Item 2	0.74			
Item 3		0.35		
Item 4				0.97
Item 5	0.75			
Item 6		0.53		
Item 7				
Item 8			0.90	
Item 9	0.47			
Item 10				
Item 11		0.70		
Item 12	0.36	0.48		
CR	0.78			
VE	0.45			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table G.11: French Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.46	
Item 2				
Item 3		0.54		
Item 4				0.37
Item 5			0.52	
Item 6			0.36	0.38
Item 7	0.65			
Item 8		0.55		
Item 9	0.70			
Item 10	0.59			
Item 11				
CR	0.76			
VE	0.34			

The main takeaway for the French EQ model exploratory factor analysis is that the results are consistent with earlier findings that all scales in this model require some attention. In particular, the Self Awareness scale is struggling. It is noted that there is an update to the development



language questionnaire. It is hoped that once the update is complete and the adaptation process is implemented for French, the changes will have a dramatic and positive impact on the French model.

G.6 French EQ Model Equating Functions

Table G.12: French-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.11	2.59	1.15	-0.21	1.07	5.04
Self Regulation	1.04	3.22	1.07	0.60	1.03	4.60
Motivation	1.10	-5.73	1.10	-4.73	1.10	-6.36
Social Awareness	1.02	4.14	1.04	2.57	0.98	7.67
Social Regulation	1.06	-0.56	1.06	-0.96	1.07	-0.10

Table G.13: French-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.18	-2.46	1.22	-4.83	1.12	1.36
Self Regulation	1.05	2.23	1.09	-0.78	1.04	3.50
Motivation	1.17	-10.59	1.18	-11.01	1.15	-10.19
Social Awareness	1.06	1.86	1.07	0.66	1.01	5.00
Social Regulation	1.07	-0.67	1.04	0.54	1.10	-2.41

G.7 French EQ Model Summary

The French version of the EQ model shows strong evidence of internal consistency reliability in four of the five scales. Self Awareness struggles in this area. Given the lower levels of internal consistency, even in the areas where in acceptable ranges, the internal structure validity requires investigation into several areas. First and foremost, it is incumbent upon the developers to ensure a proper adaptation process has been followed and that the translations are both language and culturally adequate. For example, considering the Self Awareness scale for the French-Canadien version given below in Section H.1, Table H.2 show dramatically differing results than what we see in Table G.2. It is possible that the cultural differences between Paris and Montreal account for the differing results. However, there are likely other causes to be found.

TTI SUCCESS INSIGHTS®

H EQ Analysis for French Canadien

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the French Canadien version of the TTI Success Insights EQ assessment. The study is conducted on 162 respondents. Due to the relatively small sample size, results may not be as reliable as desired at this time. Table H.1 presents the general statistics on means and standard deviations for the Australian population, as well as for a Male/Female breakdown.

Table H.1: French Canadien General Population Statistics
N = 162

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	76.38	11.00	74.52	11.66	78.25	10.05
Self Regulation	70.59	11.04	71.07	10.89	70.11	11.25
Motivation	78.65	9.23	77.54	9.03	79.75	9.35
Social Awareness	71.75	13.45	68.73	15.05	74.78	10.91
Social Regulation	70.23	11.95	69.28	12.67	71.17	11.18

H.1 Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the French Canadien version. The results are presented in Table H.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table H.2: Average French Canadien Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.37	0.20	0.21	0.36	0.28
Item 2	0.36	0.22	0.19	0.34	0.28
Item 3	0.36	0.22	0.21	0.35	0.27
Item 4	0.36	0.20	0.22	0.37	0.26
Item 5	0.35	0.21	0.20	0.35	0.26
Item 6	0.36	0.23	0.22	0.36	0.26
Item 7	0.38	0.22	0.19	0.38	0.25
Item 8	0.36	0.22	0.19	0.35	0.24
Item 9	0.37	0.22	0.19	0.35	0.26
Item 10	0.35	0.24	0.20	0.35	0.27
Item 11		0.21	0.20	0.34	0.29
Item 12		0.22	0.21	0.34	
Average	0.36	0.22	0.20	0.35	0.26
StDev	0.01	0.01	0.01	0.01	0.01

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. The scores for all scales except Motivation show acceptable levels of inter-item correlation. We have identified four items on the Motivation scale that are slightly out of the acceptable range and should be monitored.

H.2 Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the French Canadien version. The results are presented in Table H.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure

Table H.3: French Canadien Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.52	0.64	0.27	0.52	0.30
Item 2	0.58	0.33	0.53	0.62	0.34
Item 3	0.59	0.39	0.29	0.56	0.43
Item 4	0.57	0.57	0.19	0.44	0.53
Item 5	0.62	0.47	0.36	0.55	0.48
Item 6	0.56	0.26	0.22	0.51	0.47
Item 7	0.41	0.34	0.51	0.30	0.55
Item 8	0.54	0.41	0.46	0.58	0.64
Item 9	0.51	0.40	0.53	0.62	0.53
Item 10	0.62	0.17	0.40	0.58	0.44
Item 11		0.49	0.46	0.67	0.27
Item 12		0.40	0.34	0.67	
Average	0.55	0.40	0.44	0.55	0.45
StDev	0.10	0.13	0.10	0.10	0.11

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table H.3, we have highlighted potential problem areas in red. These two items are identified for monitoring and possible replacement at some point in the future. We are seeing several items on the Self Regulation, Motivation, and Social Regulation scales that are questionable. Given that we are already in the process of updating the development language questionnaire, we mark these items for monitoring with the understanding that updates intended to address such issues are already coming.

H.3 Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the French Canadien version. The results are presented in Table H.5. For convenience, Table 6.1 is reproduced as Table H.4.

Table H.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table H.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table H.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table H.5: Average French Canadien α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.85	0.76	0.73	0.86	0.79
Lower CI	0.81	0.71	0.67	0.83	0.73
Upper CI	0.88	0.82	0.79	0.89	0.83

As is presented in Table H.5, all α values for Self Awareness and Social Awareness are in the Very Good range with the remainder of the scales in the Respectable range. As noted in the intro to this appendix, the small sample size is a concern and is readily apparent in the size of the gap in the lower to upper confidence levels. Overall, all these numbers are solid and there is nothing to report based on the internal consistency values with the understanding that the sample size is likely the cause what we see in the data.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the French Canadien version of EQ has solid evidence of internal consistency reliability.

Table H.6: α-if-item-deleted for French Canadien

Est./Scale	SA	SR	MO	SoA	SoR
α	0.85	0.77	0.75	0.87	0.80
Item 1	0.84	0.73	0.75	0.86	0.80
Item 2	0.84	0.76	0.72	0.85	0.79
Item 3	0.84	0.75	0.74	0.86	0.78
Item 4	0.83	0.74	0.76	0.86	0.77
Item 5	0.83	0.75	0.74	0.86	0.78
Item 6	0.84	0.77	0.75	0.86	0.78
Item 7	0.85	0.76	0.72	0.87	0.77
Item 8	0.84	0.75	0.73	0.86	0.76
Item 9	0.84	0.75	0.72	0.85	0.77
Item 10	0.83	0.78	0.73	0.86	0.78
Item 11		0.74	0.73	0.85	0.80
Item 12		0.75	0.74	0.85	

H.4 Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

H.5 Factor Analysis

This section presents the results of the factor analysis study for the French Canadien version. The results are presented in Tables H.7 through H.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table H.7: French Canadien Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.39	
Item 2		0.62	
Item 3		0.98	
Item 4	0.62		
Item 5	0.56		
Item 6			0.98
Item 7	0.44		
Item 8	0.77		
Item 9	0.39	0.36	
Item 10	0.44		0.50
CR	0.85		
VE	0.54		

TTI SUCCESS INSIGHTS®

Some comments on how to interpret the information presented in Tables H.7 through H.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table H.8: French Canadien Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.68			
Item 2		0.99		
Item 3	0.40		0.41	
Item 4			0.94	
Item 5	0.41			
Item 6				
Item 7				
Item 8	0.47			
Item 9	0.33			
Item 10				
Item 11	0.54			
Item 12				0.98
CR	0.77			
VE	0.47			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

The main takeaway for the French Canadien EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags, but the inconsistent results are a cause for pausing some. It is the opinion of these authors that the most likely culprit for the inconsistency present is the very small sample size and all results should be viewed with a healthy skepticism until such time as a larger sample is available for analysis.

There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all EQ models at TTI Success Insights.

TTI SUCCESS INSIGHTS®

Table H.9: French Canadien Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.34
Item 2	0.94			
Item 3	0.45			
Item 4		0.36		
Item 5				0.71
Item 6				
Item 7	0.38		0.53	
Item 8			0.42	
Item 9			0.76	
Item 10	0.39			
Item 11		0.94		
Item 12				0.53
CR	0.76			
VE	0.43			

Table H.10: French Canadien Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.76
Item 2		0.80		
Item 3		0.45		
Item 4	0.49			
Item 5		0.69		
Item 6	0.62			
Item 7			0.38	
Item 8	0.46			0.47
Item 9	0.38	0.42		
Item 10			0.93	
Item 11	0.73			
Item 12	0.66			
CR	0.87			
VE	0.54			

H.6 French Canadien EQ Model Equating Functions

H.7 French Canadien EQ Model Summary

The French Canadien version of the EQ model shows some evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the French Canadien model.

Table H.11: French Candien Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2				
Item 3			0.72	
Item 4	0.53			
Item 5	0.65			0.45
Item 6	0.80			
Item 7	0.41	0.50		
Item 8	0.38		0.50	
Item 9		0.69		
Item 10		0.73		
Item 11				
CR	0.79			
VE	0.41			

Table H.12: French Canadien-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.95	4.68	0.90	8.27	1.00	-0.28
Self Regulation	1.10	-6.72	1.09	-5.20	1.09	-6.87
Motivation	1.21	-18.08	1.22	-17.24	1.20	-19.32
Social Awareness	0.86	10.63	0.79	15.61	0.98	2.20
Social Regulation	0.97	3.29	0.93	6.61	1.01	-0.63

Table H.13: French Canadien-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.06	-4.12	1.06	-3.75	1.12	-9.74
Self Regulation	1.17	-11.61	1.19	-12.24	1.15	-11.50
Motivation	1.32	-26.88	1.36	-27.92	1.32	-29.20
Social Awareness	0.93	5.32	0.86	10.23	1.08	-5.38
Social Regulation	1.05	-2.64	1.02	0.41	1.12	-8.41

I EQ Analysis for German

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the German version of the TTI Success Insights EQ assessment. The study is conducted on 8004 respondents. Table I.1 presents the general statistics on means and standard deviations for the German population, as well as for a Male/Female breakdown.

Table I.1: German General Population Statistics
N = 8004

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	77.19	9.66	75.91	9.59	78.47	9.55
Self Regulation	69.95	11.30	72.12	10.63	67.78	11.52
Motivation	76.58	10.54	77.84	10.27	75.32	10.66
Social Awareness	70.00	11.53	66.90	11.56	73.10	10.63
Social Regulation	69.88	9.86	70.10	9.96	69.67	9.76

I.1 Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the German version. The results are presented in Table I.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table I.2: Average German Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.26	0.22	0.25	0.24	0.17
Item 2	0.25	0.22	0.24	0.23	0.19
Item 3	0.24	0.23	0.26	0.24	0.18
Item 4	0.25	0.23	0.27	0.25	0.18
Item 5	0.24	0.23	0.25	0.23	0.19
Item 6	0.25	0.23	0.26	0.25	0.18
Item 7	0.25	0.24	0.25	0.26	0.17
Item 8	0.23	0.24	0.24	0.25	0.17
Item 9	0.24	0.24	0.24	0.24	0.18
Item 10	0.25	0.25	0.23	0.25	0.18
Item 11		0.23	0.25	0.23	0.20
Item 12		0.25	0.25	0.23	
Average	0.25	0.23	0.25	0.24	0.18
StDev	0.01	0.01	0.01	0.01	0.01

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. The Social Regulation



construct is of some concern. With an average inter-item correlation score that falls outside of the more restrictive limit mentioned highlights a need to explore why this has occurred. However, we note that an update is coming to the development language questionnaire that may address these concerns. We mark this scale for monitoring for the time being.

I.2 Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the German version. The results are presented in Table I.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure

Table I.3: German Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.33	0.55	0.42	0.48	0.47
Item 2	0.39	0.55	0.55	0.57	0.24
Item 3	0.47	0.42	0.31	0.43	0.35
Item 4	0.42	0.48	0.25	0.33	0.36
Item 5	0.44	0.44	0.43	0.52	0.28
Item 6	0.43	0.48	0.31	0.42	0.33
Item 7	0.39	0.36	0.44	0.23	0.46
Item 8	0.54	0.41	0.56	0.37	0.47
Item 9	0.45	0.33	0.55	0.46	0.34
Item 10	0.42	0.29	0.58	0.31	0.40
Item 11		0.50	0.42	0.57	0.18
Item 12		0.29	0.46	0.52	
Average	0.43	0.42	0.44	0.43	0.45
StDev	0.05	0.09	0.11	0.11	0.11

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table C.3, we have highlighted potential problem areas in red. In particular, we again see the Social Regulation scale needs attention and it is again marked for monitoring. There are some other areas of potential concern, but this are limited to a very small number of items that we believe will be addressed in the forthcoming update to the development language questionnaire and subsequent adaptation to the German language version.

I.3 Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the German version. The results are presented in Table I.5. For convenience, Table 6.1 is reproduced as Table I.4.

Table I.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table I.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table I.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table I.5: Average German α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.76	0.78	0.79	0.79	0.69
Lower CI	0.75	0.77	0.78	0.78	0.68
Upper CI	0.77	0.79	0.79	0.80	0.70

As is presented in Table I.5, all α values are in the Respectable range with the lone exception being the dip to 0.69 for the Social Regulation scale, implying an Undesirable internal consistency. All these numbers are solid with the exception of Social Regulation and there is nothing to report based on the internal consistency values otherwise.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the German version of EQ has solid evidence of internal consistency reliability on four of the five scales. The results of the internal consistency reliability analyses are consistent with those of the previous two sections and we again note that the Social Regulation construct is worthy of continued monitoring.

Table I.6: α-if-item-deleted for German

Est./Scale	SA	SR	MO	SoA	SoR
α	0.77	0.79	0.80	0.79	0.71
Item 1	0.76	0.76	0.79	0.78	0.67
Item 2	0.75	0.76	0.77	0.76	0.71
Item 3	0.74	0.77	0.80	0.78	0.69
Item 4	0.75	0.76	0.80	0.79	0.69
Item 5	0.74	0.77	0.79	0.77	0.70
Item 6	0.75	0.77	0.80	0.78	0.69
Item 7	0.75	0.78	0.78	0.80	0.67
Item 8	0.73	0.77	0.77	0.78	0.67
Item 9	0.74	0.78	0.77	0.77	0.69
Item 10	0.75	0.78	0.77	0.79	0.68
Item 11		0.76	0.79	0.77	0.72
Item 12		0.78	0.78	0.77	

I.4 Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

I.5 Factor Analysis

This section presents the results of the factor analysis study for the German version. The results are presented in Tables I.7 through I.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table I.7: German Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.45	
Item 2			0.98
Item 3	0.34		0.46
Item 4	0.46		
Item 5	0.39		
Item 6		0.91	
Item 7	0.52		
Item 8	0.66		
Item 9	0.48		
Item 10		0.42	
CR	0.77		
VE	0.43		

TTI SUCCESS INSIGHTS®

Some comments on how to interpret the information presented in Tables I.7 through I.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor pattern matrix, i.e., there is little or no cross-loading present.

Table I.8: German Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.50			
Item 2	0.74			
Item 3		0.58		
Item 4		0.61		
Item 5	0.38			
Item 6		0.47		
Item 7			0.41	
Item 8	0.47			0.42
Item 9	0.33			0.46
Item 10			0.42	
Item 11	0.46			
Item 12			0.51	
CR	0.79			
VE	0.35			

Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table I.9: German Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.33
Item 2	0.65			
Item 3				
Item 4				
Item 5			0.80	
Item 6				
Item 7				0.77
Item 8		0.68		
Item 9		0.56		0.38
Item 10	0.80			
Item 11		0.38		
Item 12		0.33		
CR	0.80			
VE	0.40			

TTI SUCCESS INSIGHTS®

Table I.10: German Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.43	0.54	
Item 2	0.70	0.37		
Item 3		0.46		
Item 4		0.41		
Item 5	0.74			
Item 6		0.52		
Item 7	0.41			
Item 8			0.75	
Item 9	0.42			
Item 10	0.46			
Item 11		0.56		
Item 12				0.90
CR	0.79			
VE	0.44			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions. The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table I.11: German Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.39		
Item 2				
Item 3				0.49
Item 4		0.47		
Item 5			0.95	
Item 6		0.56		
Item 7	0.60			
Item 8		0.36		0.46
Item 9	0.57			
Item 10	0.51			
Item 11				
CR	0.79			
VE	0.41			

TTI SUCCESS INSIGHTS®

The main takeaway for the German EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

I.6 German EQ Model Equating Functions

Table I.12: German-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.08	-6.21	1.10	-6.98	1.06	-4.63
Self Regulation	1.07	-4.24	1.12	-8.26	1.06	-2.60
Motivation	1.06	-4.08	1.08	-6.10	1.05	-2.93
Social Awareness	1.01	2.07	1.03	1.11	1.00	1.96
Social Regulation	1.17	-10.71	1.18	-11.89	1.16	-9.39

Table I.13: German-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.12	-10.12	1.18	-14.21	1.06	-4.95
Self Regulation	1.06	-3.76	1.09	-5.92	1.08	-3.90
Motivation	1.10	-7.60	1.13	-10.83	1.09	-5.96
Social Awareness	0.98	3.78	1.02	1.60	0.96	5.12
Social Regulation	1.13	-7.50	1.16	-10.71	1.11	-0.60

I.7 German EQ Model Summary

The German version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the German model. In particular, we expect the updates to the development language questionnaire to have a major impact on the Social Regulation scale of the German model.

J EQ Analysis for Hungarian

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Hungarian version of the TTI Success Insights EQ assessment. The study is conducted on 342 respondents. The relatively small sample size may cause some results to be inconsistent at this time. Table J.1 presents the general statistics on means and standard deviations for the Hungarian population, as well as for a Male/Female breakdown.

Table J.1: Hungarian General Population Statistics
N = 342

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	70.51	8.69	68.34	8.19	72.69	8.66
Self Regulation	69.25	12.85	71.63	12.31	66.88	12.98
Motivation	80.15	10.87	80.58	10.98	79.71	10.78
Social Awareness	69.44	11.54	65.67	11.57	73.2	10.22
Social Regulation	70.23	10.96	70.36	10.83	70.12	11.12

J.1 Inter Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Hungarian version. The results are presented in Table J.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table J.2: Average Hungarian Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.14	0.30	0.28	0.22	0.19
Item 2	0.13	0.31	0.27	0.19	0.24
Item 3	0.13	0.33	0.29	0.21	0.21
Item 4	0.18	0.31	0.30	0.22	0.24
Item 5	0.18	0.33	0.27	0.22	0.21
Item 6	0.13	0.31	0.31	0.22	0.21
Item 7	0.15	0.35	0.29	0.24	0.21
Item 8	0.18	0.33	0.28	0.21	0.20
Item 9	0.15	0.32	0.26	0.22	0.21
Item 10	0.13	0.33	0.27	0.23	0.21
Item 11		0.31	0.28	0.20	0.23
Item 12		0.33	0.28	0.21	
Average	0.15	0.32	0.28	0.22	0.21
StDev	0.02	0.01	0.01	0.01	0.02

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Hungarian version based on inter-item correlation for the Self Regulation, Motivation, Social Awareness, and Social Regulation scales. The entire Self Awareness scale is marked for monitoring. It is noted that changes to the development language questionnaire are forthcoming and may result in improved performance of this scale.

J.2 Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Hungarian version. The results are presented in Table J.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure

Table J.3: Hungarian Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.41	0.70	0.45	0.37	0.58
Item 2	0.46	0.57	0.60	0.63	0.16
Item 3	0.39	0.47	0.43	0.46	0.46
Item 4	0.07	0.62	0.31	0.35	0.16
Item 5	0.09	0.47	0.56	0.32	0.42
Item 6	0.45	0.66	0.20	0.38	0.45
Item 7	0.31	0.31	0.39	0.13	0.41
Item 8	0.00	0.41	0.53	0.43	0.57
Item 9	0.29	0.54	0.68	0.39	0.40
Item 10	0.42	0.44	0.60	0.27	0.46
Item 11		0.62	0.53	0.61	0.29
Item 12		0.44	0.46	0.51	
Average	0.29	0.52	0.48	0.40	0.40
StDev	0.17	0.12	0.13	0.14	0.14

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table J.3, we have highlighted potential problem areas in red. We once again see some issues in the Self Awareness scale as well as a smattering of other items showing inconsistencies. All items in red or orange in the table are marked for monitoring. We again not the upcoming modifications to the development language questionnaire and subsequent adaptations to follow.

J.3 Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Hungarian version. The results are presented in Table J.5. For convenience, Table 6.1 is reproduced as Table J.4.

Table J.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table J.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table J.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table J.5: Average Hungarian α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.57	0.85	0.82	0.76	0.74
Lower CI	0.50	0.83	0.79	0.72	0.70
Upper CI	0.63	0.87	0.85	0.89	0.78

As is presented in Table J.5, all α values are in the Very Good or Respectable ranges with the lone exception being the Self Awareness scale which has fallen into the Unacceptable range. It appears necessary to take a strong look at the Hungarian Self Awareness scale.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are several items identified as having an increase in the value of α at this time. We conclude that the Hungarian version of EQ has some evidence of internal consistency reliability with the Self Awareness scale being identified as problematic at this time. We do note the relatively small sample size studied for this analysis.

Table J.6: α-if-item-deleted for Hungarian

Est./Scale	SA	SR	MO	SoA	SoR
α	0.64	0.85	0.82	0.77	0.75
Item 1	0.58	0.83	0.81	0.76	0.70
Item 2	0.57	0.83	0.80	0.73	0.76
Item 3	0.58	0.84	0.81	0.75	0.72
Item 4	0.66	0.83	0.82	0.76	0.76
Item 5	0.66	0.84	0.80	0.76	0.73
Item 6	0.57	0.83	0.83	0.75	0.72
Item 7	0.61	0.85	0.82	0.78	0.73
Item 8	0.67	0.85	0.81	0.75	0.71
Item 9	0.61	0.84	0.79	0.75	0.73
Item 10	0.58	0.84	0.80	0.77	0.72
Item 11		0.83	0.81	0.73	0.75
Item 12		0.84	0.81	0.74	

J.4 Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

J.5 Factor Analysis

This section presents the results of the factor analysis study for the Hungarian version. The results are presented in Tables J.7 through J.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table J.7: Hungarian Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1	0.51		
Item 2		0.80	
Item 3		0.62	
Item 4			0.48
Item 5			0.82
Item 6	0.76		
Item 7	0.36		
Item 8			
Item 9		0.36	
Item 10	0.54		
CR	0.64		
VE	0.38		

Some comments on how to interpret the information presented in Tables J.7 through J.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor pattern



matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table J.8: Hungarian Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.74	0.36		
Item 2	0.48		0.38	
Item 3		0.50		
Item 4		0.94		
Item 5			0.43	
Item 6	0.35	0.43	0.46	
Item 7				0.68
Item 8			0.43	
Item 9	0.35		0.46	
Item 10			0.56	
Item 11	0.68			
Item 12			0.38	
CR	0.85			
VE	0.49			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table J.9: Hungarian Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.95	
Item 2		0.91		
Item 3	0.34			
Item 4	0.34			
Item 5	0.47			
Item 6				0.56
Item 7	0.34			
Item 8	0.60			
Item 9	0.68			
Item 10	0.46	0.38		
Item 11	0.64			
Item 12	0.43			
CR	0.83			
VE	0.46			

TTI SUCCESS INSIGHTS®

Table J.10: Hungarian Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.63
Item 2		0.64		
Item 3	0.46			
Item 4			0.96	
Item 5		0.44		
Item 6	0.58			
Item 7				
Item 8				0.57
Item 9		0.62		
Item 10		0.34		
Item 11	0.74			
Item 12	0.40	0.44		
CR	0.78			
VE	0.43			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table J.11: Hungarian Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.35	0.42	
Item 2				0.34
Item 3		0.54		0.49
Item 4				
Item 5			0.95	
Item 6				0.64
Item 7	0.47			
Item 8		0.87		
Item 9	0.73			
Item 10	0.66			
Item 11				
CR	0.79			
VE	0.41			

The main takeaway for the Hungarian EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good with the exception of the Self Awareness scale. There is some lack of balance and some cross-loading that we have identified and we shall work, over time,



to rectify this situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

J.6 Hungary EQ Model Equating Functions

Table J.12: Hungary-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.20	-7.49	1.28	-12.28	1.17	-6.41
Self Regulation	0.94	5.48	0.96	3.20	0.94	6.36
Motivation	1.03	-5.28	1.01	-3.44	1.04	-6.61
Social Awareness	1.01	2.70	1.03	2.43	1.04	-1.08
Social Regulation	1.06	-2.85	1.09	-5.51	1.02	0.05

Table J.13: Hungary-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.28	-13.88	1.40	-20.85	1.25	-12.51
Self Regulation	1.00	1.68	1.03	-1.45	1.00	2.45
Motivation	1.10	-11.59	1.10	-11.43	1.12	-13.54
Social Awareness	1.07	-2.29	1.11	-3.31	1.12	-6.76
Social Regulation	1.13	-8.52	1.15	-10.27	1.10	-5.14

J.7 Hungarian EQ Model Summary

The Hungarian version of the EQ model shows some evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Hungarian model.

K EQ Analysis for Italian

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Italian version of the TTI Success Insights EQ assessment. The study is conducted on 848 respondents. Due to the relatively small number of respondents, some results may be inconsistent. Table K.1 presents the general statistics on means and standard deviations for the Australian population, as well as for a Male/Female breakdown.

Table K.1: Italian General Population Statistics
N = 848

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	73.47	12.16	73.51	11.50	73.43	12.81
Self Regulation	65.79	13.00	68.70	11.76	62.89	13.54
Motivation	73.32	12.55	74.95	12.76	71.68	12.28
Social Awareness	69.24	12.43	67.25	11.80	71.23	11.74
Social Regulation	66.56	11.74	66.87	11.67	66.26	11.81

K.1 Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Italian version. The results are presented in Table K.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table K.2: Average Italian Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.33	0.26	0.28	0.24	0.21
Item 2	0.35	0.29	0.27	0.23	0.23
Item 3	0.33	0.29	0.30	0.26	0.22
Item 4	0.33	0.27	0.31	0.25	0.22
Item 5	0.31	0.28	0.28	0.24	0.21
Item 6	0.33	0.27	0.31	0.24	0.22
Item 7	0.33	0.39	0.28	0.26	0.21
Item 8	0.32	0.28	0.28	0.24	0.21
Item 9	0.32	0.30	0.27	0.24	0.23
Item 10	0.32	0.29	0.27	0.25	0.21
Item 11		0.27	0.28	0.22	0.24
Item 12		0.30	0.28	0.23	
Average	0.33	0.28	0.28	0.24	0.22
StDev	0.01	0.01	0.01	0.01	0.01

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Italian version based on inter-item correlation.

K.2 Item-Total Correlation Analysis

This section presents the results of the item-total correlation study for the Italian version. The results are presented in Table K.3. As is discussed in Section 11, corrected item-total correlation serves as a proxy for evidence of validity based on internal structure

Table K.3: Italian Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.49	0.67	0.52	0.48	0.49
Item 2	0.40	0.43	0.60	0.58	0.27
Item 3	0.48	0.44	0.37	0.30	0.40
Item 4	0.50	0.57	0.23	0.33	0.38
Item 5	0.61	0.48	0.49	0.48	0.53
Item 6	0.48	0.60	0.25	0.42	0.41
Item 7	0.50	0.39	0.47	0.23	0.45
Item 8	0.60	0.49	0.55	0.43	0.51
Item 9	0.54	0.36	0.65	0.46	0.32
Item 10	0.55	0.39	0.60	0.30	0.47
Item 11		0.58	0.49	0.61	0.20
Item 12		0.37	0.54	0.57	
Average	0.52	0.48	0.48	0.43	0.38
StDev	0.06	0.10	0.13	0.12	0.01

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table K.3, we have highlighted potential problem areas in red. These items belong to the Motivation, Social Awareness, and Social Regulation constructs. Several of the identified items have item-to-total correlations approach 0.20, well below our accepted lower bound of 0.30. We mark this items for monitoring and note that there is an upcoming update to the development language version of the assessment that is likely to have a dramatic and positive impact on the Italian version after the adaptation process.

K.3 Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Italian language. The results are presented in Table K.5. For convenience, Table 6.1 is reproduced as Table K.4.

Table K.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table K.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table K.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table K.5: Average Italian α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.83	0.82	0.81	0.79	0.75
Lower CI	0.81	0.80	0.80	0.77	0.72
Upper CI	0.84	0.84	0.83	0.81	0.77

As is presented in Table K.5, α values are in the Very Good ranges for the Self Awareness, Self Regulation. and Motivation scales. The values for Social Awareness and Social Regulation dip below 0.80 implying a Respectable internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Italian version of EQ has reasonable evidence of internal consistency reliability.

Table K.6: α-if-item-deleted for Italian

Est./Scale	SA	SR	MO	SoA	SoR
α	0.83	0.84	0.83	0.79	0.76
Item 1	0.82	0.80	0.81	0.77	0.73
Item 2	0.83	0.82	0.80	0.76	0.75
Item 3	0.82	0.82	0.82	0.79	0.74
Item 4	0.82	0.81	0.83	0.79	0.74
Item 5	0.80	0.81	0.81	0.77	0.72
Item 6	0.82	0.80	0.83	0.78	0.74
Item 7	0.82	0.82	0.81	0.80	0.73
Item 8	0.81	0.81	0.81	0.78	0.72
Item 9	0.81	0.82	0.80	0.77	0.75
Item 10	0.81	0.82	0.80	0.79	0.73
Item 11		0.80	0.81	0.76	0.76
Item 12		0.82	0.81	0.76	

K.4 Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

K.5 Factor Analysis

This section presents the results of the factor analysis study for the Italian version. The results are presented in Tables K.7 through K.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table K.7: Italian Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.58	
Item 2			0.61
Item 3			0.68
Item 4	0.59		
Item 5	0.58		
Item 6		0.76	
Item 7	0.58		
Item 8	0.71		
Item 9	0.54		
Item 10	0.45	0.37	
CR	0.83		
VE	0.46		

Some comments on how to interpret the information presented in Tables K.7 through K.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table K.8: Italian Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.38			0.67
Item 2			0.34	0.34
Item 3	0.58			
Item 4	0.78			
Item 5			0.83	
Item 6	0.48			
Item 7		0.53		
Item 8			0.33	
Item 9		0.32		
Item 10		0.42		
Item 11				0.51
Item 12		0.52		
CR	0.83			
VE	0.43			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table K.9: Italian Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.49		
Item 2	0.67			
Item 3	0.38			
Item 4				
Item 5		0.78		
Item 6				
Item 7			0.57	
Item 8				0.81
Item 9			0.68	
Item 10	0.69			
Item 11				0.38
Item 12		0.55		
CR	0.83			
VE	0.44			

TTI SUCCESS INSIGHTS®

Table K.10: Italian Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.89	
Item 2		0.63		
Item 3	0.41			
Item 4	0.36			
Item 5		0.53		
Item 6	0.48			
Item 7				0.57
Item 8	0.35		0.45	
Item 9		0.58		
Item 10		0.36		0.42
Item 11	0.74			
Item 12	0.48			
CR	0.80			
VE	0.42			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table K.11: Italian Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.49			
Item 2				0.37
Item 3	0.57			0.41
Item 4			0.51	
Item 5	0.57			
Item 6			0.40	
Item 7		0.57		
Item 8	0.54			
Item 9				
Item 10		0.62		
Item 11				
CR	0.76			
VE	0.34			

The main takeaway for the Italian EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation

and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

K.6 Italy EQ Model Equating Functions

Table K.12: Italy-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.86	14.05	0.91	8.25	0.79	20.44
Self Regulation	0.93	9.46	1.01	2.93	0.91	12.59
Motivation	0.89	11.79	0.91	9.49	0.88	13.29
Social Awareness	0.95	7.10	0.96	5.59	0.90	10.92
Social Regulation	0.99	5.69	1.01	3.52	0.96	7.94

Table K.13: Italy-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.91	9.16	1.00	1.81	0.83	16.79
Self Regulation	0.98	6.14	1.06	-0.91	0.95	9.69
Motivation	0.95	6.74	0.98	3.69	0.94	8.38
Social Awareness	0.99	3.88	1.01	2.02	0.97	6.16
Social Regulation	1.04	2.25	1.03	1.77	1.01	3.89

K.7 Italian EQ Model Summary

The Italian version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Italian model.

L EQ Analysis for Portuguese

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Portuguese version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 180 respondents. Table L.1 presents the general statistics on means and standard deviations for the Portuguese population, as well as for a Male/Female breakdown.

Table L.1: Portuguese General Population Statistics
N = 180

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	76.70	10.09	75.64	10.73	77.76	9.35
Self Regulation	69.47	10.99	72.00	10.53	66.93	10.92
Motivation	75.54	11.04	77.60	11.08	73.48	10.65
Social Awareness	70.77	10.27	68.16	9.81	73.39	10.10
Social Regulation	69.29	10.79	70.89	11.01	68.21	10.51

L.1 Portuguese Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Portuguese version. The results are presented in Table L.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table L.2: Average Portuguese Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.32	0.25	0.27	0.22	0.25
Item 2	0.32	0.26	0.25	0.19	0.24
Item 3	0.31	0.28	0.27	0.22	0.24
Item 4	0.31	0.27	0.28	0.21	0.25
Item 5	0.30	0.27	0.27	0.19	0.23
Item 6	0.31	0.27	0.29	0.21	0.25
Item 7	0.36	0.27	0.26	0.21	0.22
Item 8	0.30	0.27	0.26	0.22	0.23
Item 9	0.33	0.27	0.26	0.20	0.22
Item 10	0.31	0.27	0.27	0.21	0.23
Item 11		0.25	0.28	0.20	0.26
Item 12		0.28	0.26	0.19	
Mean	0.32	0.27	0.27	0.20	0.24
StDev	0.01	0.01	0.01	0.01	0.01

TTI SUCCESS INSIGHTS®

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Portuguese version based on inter-item correlation in four of the five constructs. The average score for Social Awareness at 0.20, combined with multiple items scoring less than 0.20 implies this scale needs to be monitored.

L.2 Portuguese Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Portuguese version. The results are presented in Table L.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table L.3: Portuguese Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.48	0.62	0.47	0.23	0.33
Item 2	0.49	0.55	0.64	0.51	0.35
Item 3	0.59	0.33	0.40	0.27	0.46
Item 4	0.57	0.42	0.33	0.29	0.34
Item 5	0.61	0.46	0.41	0.56	0.47
Item 6	0.57	0.39	0.24	0.36	0.35
Item 7	0.22	0.43	0.56	0.35	0.60
Item 8	0.60	0.46	0.56	0.29	0.48
Item 9	0.38	0.44	0.55	0.48	0.56
Item 10	0.52	0.45	0.49	0.34	0.46
Item 11		0.65	0.34	0.48	0.26
Item 12		0.30	0.52	0.51	
Mean	0.50	0.46	0.46	0.39	0.42
StDev	0.12	0.10	0.12	0.11	0.11

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table L.3, we have highlighted potential problem areas in red. These items are identified for monitoring and possible replacement at some point in the future. In particular, the Social Awareness results confirm what is seen in the previous section. However, there is no alarm to be sounded at this time. The mean item-to-total correlation scores are are in an acceptable range and the standard deviations show consistent scoring.

L.3 Portuguese Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Portuguese version. The results are presented in Table L.5. For convenience, Table 6.1 is reproduced as Table L.4.

Table L.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table L.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table L.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table L.5: Average Portuguese α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.81	0.76	0.79	0.75	0.76
Lower CI	0.76	0.80	0.75	0.69	0.71
Upper CI	0.85	0.84	0.84	0.80	0.81

As is presented in Table L.5, all α values are in the Respectable range with the lone exception being the rise to 0.81 for the Self Awareness scale, implying a Very Good internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Portuguese version of EQ has solid evidence of internal consistency reliability.

Table L.6: α-if-item-deleted for Portuguese

Est./Scale	SA	SR	MO	SoA	SoR
α	0.82	0.81	0.82	0.76	0.77
Item 1	0.81	0.78	0.80	0.76	0.77
Item 2	0.81	0.79	0.79	0.72	0.76
Item 3	0.80	0.81	0.81	0.75	0.75
Item 4	0.80	0.81	0.81	0.75	0.77
Item 5	0.80	0.80	0.81	0.72	0.75
Item 6	0.80	0.81	0.82	0.74	0.76
Item 7	0.83	0.80	0.79	0.74	0.73
Item 8	0.80	0.80	0.79	0.75	0.75
Item 9	0.82	0.80	0.79	0.73	0.74
Item 10	0.80	0.81	0.80	0.74	0.75
Item 11		0.78	0.81	0.73	0.78
Item 12		0.81	0.80	0.72	

L.4 Portuguese Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

L.5 Portuguese Factor Analysis

This section presents the results of the factor analysis study for the Portuguese version. The results are presented in Tables L.7 through L.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table L.7: Portuguese Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1			0.49
Item 2		0.66	
Item 3		0.84	
Item 4	0.57		
Item 5	0.55		
Item 6			0.82
Item 7			
Item 8	0.75		
Item 9			
Item 10			0.50
CR	0.83		
VE	0.47		

Some comments on how to interpret the information presented in Tables L.7 through L.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table L.8: Portuguese Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				0.86
Item 2		0.72		
Item 3			0.48	
Item 4			0.79	
Item 5		0.43		
Item 6				
Item 7	0.64			
Item 8	0.38	0.43		
Item 9		0.50		
Item 10	0.46			
Item 11	0.53			
Item 12	0.46			
CR	0.84			
VE	0.45			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table L.9: Portuguese Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.40			
Item 2		0.74		
Item 3		0.42	0.48	
Item 4				
Item 5	0.54			
Item 6			0.47	
Item 7	0.66			
Item 8	0.44			0.50
Item 9	0.69			0.44
Item 10		0.69		
Item 11				
Item 12	0.72			
CR	0.85			
VE	0.44			

TTI SUCCESS INSIGHTS®

Table L.10: Portuguese Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.47	
Item 2	0.44			0.88
Item 3		0.36		
Item 4				
Item 5	0.71			
Item 6		0.50		
Item 7	0.57			
Item 8			0.80	
Item 9		0.37		
Item 10	0.62			
Item 11		0.83		
Item 12	0.39	0.37		
CR	0.78			
VE	0.45			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table L.11: Portuguese Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.36		
Item 2				
Item 3				0.63
Item 4	0.43			
Item 5		0.66		
Item 6		0.37		
Item 7	0.68	0.46		
Item 8				0.41
Item 9	0.65			
Item 10	0.72			
Item 11			0.98	
CR	0.76			
VE	0.45			

The main takeaway for the Portuguese EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no major red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this



situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all EQ models at TTI Success Insights.

L.6 Portuguese EQ Model Equating Functions

Table L.12: Portuguese-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.03	-2.17	0.98	1.34	1.08	-5.63
Self Regulation	1.10	-5.84	1.13	-8.89	1.12	-5.61
Motivation	1.01	0.64	1.00	0.26	1.05	-1.06
Social Awareness	1.13	-7.44	1.21	-12.68	1.06	-2.19
Social Regulation	1.07	-3.01	1.07	-4.35	1.08	-2.04

Table L.13: Portuguese-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.14	-11.05	1.11	-9.19	1.18	-13.97
Self Regulation	1.16	-9.99	1.20	-14.62	1.18	-9.89
Motivation	1.11	-7.18	1.13	-10.34	1.15	-8.93
Social Awareness	1.19	-12.15	1.30	-19.20	1.14	-8.27
Social Regulation	1.14	-8.25	1.15	-10.13	1.15	-7.22

L.7 Portuguese EQ Model Summary

The Portuguese version of the EQ model shows reasonable evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Portuguese model.

M EQ Analysis for Spain

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Spanish (Spain) (Spanish ES) version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 2284 respondents. Table M.1 presents the general statistics on means and standard deviations for the Spanish population, as well as for a Male/Female breakdown.

Table M.1: Spanish ES General Population Statistics
N = 2284

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	75.90	10.78	74.63	11.12	77.16	10.28
Self Regulation	67.54	12.48	69.70	12.01	65.38	12.57
Motivation	75.36	10.98	75.59	11.09	75.12	10.86
Social Awareness	72.31	11.76	68.92	11.98	75.70	10.50
Social Regulation	69.89	10.83	69.33	11.20	70.46	10.41

M.1 Spanish ES Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Spanish ES version. The results are presented in Table M.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table M.2: Average Spanish ES Inter-item Correlations
TTI SI Emotional Quotient

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.32	0.23	0.24	0.24	0.20
Item 2	0.31	0.25	0.24	0.22	0.22
Item 3	0.30	0.25	0.24	0.24	0.20
Item 4	0.30	0.24	0.26	0.25	0.21
Item 5	0.29	0.25	0.24	0.23	0.19
Item 6	0.31	0.25	0.25	0.24	0.19
Item 7	0.30	0.27	0.24	0.26	0.19
Item 8	0.29	0.25	0.23	0.25	0.19
Item 9	0.31	0.24	0.23	0.23	0.20
Item 10	0.31	0.26	0.23	0.25	0.20
Item 11		0.24	0.24	0.23	0.23
Item 12		0.26	0.23	0.23	
Mean	0.31	0.25	0.24	0.24	0.20
StDev	0.01	0.01	0.01	0.01	0.01

TTI SUCCESS INSIGHTS®

The results are quite solid overall except for the Social Regulation scale. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. With the Social Regulation scale showing multiple items under the lower acceptable level of 0.20 and the average inter-item correlation at the 0.20 level, the entire scale is marked for monitoring.

M.2 Spanish ES Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Spanish version. The results are presented in Table M.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table M.3: Spanish ES Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.38	0.59	0.43	0.42	0.48
Item 2	0.45	0.47	0.42	0.61	0.28
Item 3	0.52	0.43	0.39	0.42	0.48
Item 4	0.51	0.52	0.23	0.37	0.37
Item 5	0.59	0.47	0.39	0.49	0.56
Item 6	0.44	0.41	0.28	0.46	0.54
Item 7	0.52	0.27	0.44	0.21	0.54
Item 8	0.57	0.44	0.50	0.38	0.58
Item 9	0.48	0.54	0.54	0.49	0.47
Item 10	0.46	0.32	0.54	0.33	0.47
Item 11		0.51	0.42	0.50	0.20
Item 12		0.36	0.51	0.54	
Mean	0.49	0.44	0.43	0.43	0.45
StDev	0.06	0.09	0.10	0.11	0.12

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table M.3, we have highlighted potential problem areas in red. These items are identified for monitoring and possible replacement at some point in the future. There is no alarm to be sounded at this time. The mean item-to-total correlation scores are are in an acceptable range and the standard deviations show consistent scoring.

M.3 Spanish ES Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Spanish (Spain) language. The results are presented in Table M.5. For convenience, Table 6.1 is reproduced as Table M.4.

Table M.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table M.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table M.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table M.5: Average Spanish ES α Coefficient Data TTI SI Emotional Quotient

Est./Scale	SoA	MO	SA	SR	SoR
α	0.81	0.80	0.77	0.79	0.72
Lower CI	0.80	0.79	0.76	0.78	0.71
Upper CI	0.82	0.81	0.79	0.80	0.74

As is presented in Table M.5, all α values are in the Very Good ranges with the lone exception being the dip to 0.79 for the Social Regulation scale, implying a Respectable internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Spanish ES version of EQ has solid evidence of internal consistency reliability.

Table M.6: α-if-item-deleted for Spanish ES

Est./Scale	SA	SR	MO	SoA	SoR
α	0.81	0.80	0.79	0.79	0.74
Item 1	0.81	0.77	0.77	0.78	0.71
Item 2	0.80	0.78	0.77	0.76	0.74
Item 3	0.80	0.79	0.78	0.78	0.72
Item 4	0.80	0.78	0.79	0.78	0.73
Item 5	0.79	0.78	0.78	0.77	0.71
Item 6	0.80	0.79	0.79	0.78	0.71
Item 7	0.79	0.80	0.77	0.80	0.71
Item 8	0.79	0.78	0.77	0.78	0.70
Item 9	0.80	0.77	0.76	0.77	0.72
Item 10	0.80	0.80	0.76	0.79	0.72
Item 11		0.78	0.77	0.77	0.75
Item 12		0.79	0.76	0.77	

M.4 Spanish ES Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

M.5 Spanish ES Factor Analysis

This section presents the results of the factor analysis study for the Spanish version. The results are presented in Tables M.7 through M.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table M.7: Spanish ES Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1			0.39
Item 2		0.52	
Item 3		0.93	
Item 4	0.59		
Item 5	0.53		
Item 6			0.76
Item 7	0.56		
Item 8	0.69		
Item 9	0.47		
Item 10			0.58
CR	0.82		
VE	0.45		

Some comments on how to interpret the information presented in Tables M.7 through M.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor.

Table M.8: Spanish ES Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.73			
Item 2	0.40			0.46
Item 3		0.59		
Item 4		0.65		
Item 5				0.69
Item 6				
Item 7			0.35	
Item 8			0.38	
Item 9			0.37	
Item 10			0.46	
Item 11	0.57			
Item 12			0.55	
CR	0.83			
VE	0.38			

Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table M.9: Spanish ES Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.41	
Item 2		0.67		
Item 3		0.42		
Item 4				
Item 5			0.70	
Item 6				0.48
Item 7	0.43			0.37
Item 8	0.45			
Item 9	0.70			
Item 10		0.63		
Item 11	0.39			
Item 12	0.40		0.53	
CR	0.82			
VE	0.37			

TTI SUCCESS INSIGHTS®

Table M.10: Spanish ES Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.45	
Item 2		0.75		
Item 3	0.46			
Item 4	0.41			
Item 5		0.63		
Item 6	0.59			
Item 7				
Item 8			0.97	
Item 9		0.54		
Item 10				0.68
Item 11	0.66			
Item 12	0.45			
CR	0.82			
VE	0.44			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions. The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table M.11: Spanish ES Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.48		
Item 2				
Item 3		0.38		
Item 4			0.53	
Item 5		0.68		
Item 6			0.46	
Item 7	0.55			
Item 8				0.72
Item 9	0.59			
Item 10	0.62			
Item 11				
CR	0.76			
VE	0.35			

The main takeaway for the Spanish ES EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance



and some cross-loading that we have identified and we shall work, over time, to rectify this situation and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all EQ models at TTI Success Insights.

M.6 Spanish ES EQ Model Equating Functions

Table M.12: Spanish ES-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.96	3.67	0.95	4.90	0.98	2.55
Self Regulation	0.97	5.20	0.99	3.38	0.98	5.76
Motivation	1.02	0.40	1.00	2.33	1.03	-1.26
Social Awareness	0.99	1.17	0.99	1.52	1.02	-1.58
Social Regulation	1.07	-3.38	1.05	-1.90	1.09	-5.19

Table M.13: Spanish ES-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.05	-3.35	1.03	-1.85	1.04	-2.15
Self Regulation	1.00	2.91	1.02	1.45	1.02	3.03
Motivation	1.09	-5.12	1.08	-4.37	1.09	-5.06
Social Awareness	1.02	-1.25	1.03	-1.57	1.09	-7.75
Social Regulation	1.11	-6.37	1.08	-3.89	1.14	-8.86

M.7 Spanish ES EQ Model Summary

The Spanish ES version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Spanish ES model.

N EQ Analysis for Spanish North America

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Spanish North America[6] (Spanish NA) version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 2818 respondents. Table N.1 presents the general statistics on means and standard deviations for the Spanish NA population, as well as for a Male/Female breakdown.

Table N.1: Spanish NA General Population Statistics
N = 2818

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	78.24	11.35	78.19	11.11	78.28	11.58
Self Regulation	70.73	12.67	72.67	11.69	68.80	13.31
Motivation	80.25	11.52	81.52	10.69	78.98	12.17
Social Awareness	72.58	10.73	70.64	10.58	74.51	10.53
Social Regulation	71.95	11.61	72.25	11.42	71.66	11.78

N.1 Spanish NA Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Spanish NA version. The results are presented in Table N.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values for four of the scales are well within these limits, there are no issues identified for the Spanish NA version based on inter-item correlation for those scales. However, the entire Social Awareness construct is marked for monitoring.

N.2 Spanish NA Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Spanish NA version. The results are presented in Table N.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some

[6]Spanish North America contains all Spanish speaking respondents in the western hemisphere.

TTI SUCCESS INSIGHTS®

Table N.2: Average Spanish NA Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.35	0.25	0.30	0.19	0.24
Item 2	0.35	0.27	0.30	0.17	0.24
Item 3	0.34	0.28	0.30	0.20	0.24
Item 4	0.34	0.28	0.33	0.20	0.24
Item 5	0.32	0.27	0.31	0.17	0.25
Item 6	0.33	0.28	0.32	0.20	0.23
Item 7	0.33	0.27	0.31	0.20	0.23
Item 8	0.33	0.28	0.30	0.19	0.22
Item 9	0.34	0.26	0.30	0.18	0.23
Item 10	0.33	0.28	0.30	0.19	0.22
Item 11		0.26	0.31	0.19	0.26
Item 12		0.27	0.29	0.18	
Mean	0.34	0.27	0.31	0.19	0.24
StDev	0.01	0.01	0.01	0.01	0.01

Table N.3: Spanish Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.45	0.63	0.53	0.33	0.40
Item 2	0.44	0.50	0.57	0.52	0.37
Item 3	0.51	0.42	0.53	0.27	0.43
Item 4	0.48	0.44	0.27	0.23	0.32
Item 5	0.61	0.48	0.45	0.51	0.44
Item 6	0.55	0.37	0.42	0.31	0.44
Item 7	0.54	0.44	0.50	0.27	0.47
Item 8	0.59	0.37	0.52	0.39	0.54
Item 9	0.48	0.53	0.57	0.43	0.44
Item 10	0.54	0.37	0.60	0.36	0.54
Item 11		0.54	0.48	0.32	0.23
Item 12		0.48	0.62	0.48	
Mean	0.52	0.46	0.50	0.37	0.42
StDev	0.06	0.08	0.09	0.10	0.09

caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table N.3, we have highlighted potential problem areas in red. These items are identified for monitoring and possible replacement at some point in the future. The Social Awareness scale shows some need for attention. The mean item-to-total correlation scores are are in an acceptable range

and the standard deviations show consistent scoring.

N.3 Spanish NA Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Spanish NA version. The results are presented in Table N.5. For convenience, Table 6.1 is reproduced as Table N.4.

Table N.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table N.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table N.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table N.5: Average Spanish NA α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.82	0.81	0.83	0.73	0.76
Lower CI	0.81	0.80	0.82	0.71	0.75
Upper CI	0.83	0.82	0.84	0.74	0.77

As is presented in Table N.5, all α values are in the Very Good ranges with the lone exception being the dip to 0.79 for the Social Regulation scale, implying a Respectable internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Spanish NA version of EQ has some evidence of internal consistency reliability.

Table N.6: α-if-item-deleted for Spanish NA

Est./Scale	SA	SR	MO	SoA	SoR
α	0.84	0.82	0.84	0.74	0.77
Item 1	0.83	0.79	0.83	0.73	0.76
Item 2	0.83	0.80	0.83	0.70	0.76
Item 3	0.82	0.81	0.83	0.73	0.76
Item 4	0.82	0.81	0.85	0.74	0.77
Item 5	0.81	0.80	0.83	0.70	0.75
Item 6	0.82	0.81	0.84	0.73	0.75
Item 7	0.82	0.80	0.83	0.73	0.75
Item 8	0.81	0.81	0.83	0.73	0.74
Item 9	0.82	0.80	0.82	0.71	0.75
Item 10	0.82	0.81	0.82	0.72	0.74
Item 11		0.80	0.83	0.73	0.78
Item 12		0.80	0.83	0.70	

N.4 Spanish NA Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

N.5 Spanish NA Factor Analysis

This section presents the results of the factor analysis study for the Spanish NA version. The results are presented in Tables N.7 through N.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table N.7: Spanish NA Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1		0.44	
Item 2			0.76
Item 3			0.76
Item 4	0.55		
Item 5	0.56		
Item 6		0.79	
Item 7	0.57		
Item 8	0.67		
Item 9	0.48		
Item 10	0.45	0.48	
CR	0.84		
VE	0.48		

Some comments on how to interpret the information presented in Tables N.7 through N.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor

TTI SUCCESS INSIGHTS®

pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table N.8: Spanish NA Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.69			
Item 2	0.60			
Item 3		0.54		
Item 4		0.66		
Item 5		0.36		
Item 6		0.37		
Item 7			0.37	
Item 8				0.96
Item 9	0.46			
Item 10			0.44	
Item 11	0.61			
Item 12			0.75	
CR	0.84			
VE	0.44			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table N.9: Spanish NA Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.43		
Item 2	0.73			
Item 3	0.46			
Item 4				
Item 5				0.77
Item 6				
Item 7		0.37	0.37	
Item 8			0.54	
Item 9		0.37	0.51	
Item 10	0.62			
Item 11				
Item 12		0.61		
CR	0.86			
VE	0.41			

TTI SUCCESS INSIGHTS®

Table N.10: Spanish NA Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.60		
Item 2	0.61			
Item 3				
Item 4				0.48
Item 5	0.75			
Item 6				0.52
Item 7	0.49			
Item 8		0.79		
Item 9	0.50			
Item 10	0.55			
Item 11			0.55	
Item 12	0.44			
CR	0.77			
VE	0.37			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table N.11: Spanish NA Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.37	
Item 2	0.37			
Item 3		0.53		
Item 4				
Item 5			0.53	
Item 6			0.41	
Item 7	0.60			
Item 8		0.66		
Item 9	0.35			0.67
Item 10	0.56			
Item 11				
CR	0.80			
VE	0.35			

The main takeaway for the Spanish NA EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation

and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

N.6 Spanish NA EQ Model Equating Functions

Table N.12: Spanish NA-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.92	5.20	0.95	1.46	0.87	10.08
Self Regulation	0.96	3.13	1.02	-1.52	0.92	6.16
Motivation	0.97	-0.73	1.03	-6.62	0.92	3.52
Social Awareness	1.08	-5.97	1.12	-9.46	1.01	-0.16
Social Regulation	1.00	-0.41	1.03	-3.51	0.96	2.58

Table N.13: Spanish NA-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.96	1.97	0.98	-1.46	0.92	0.64
Self Regulation	0.98	1.29	1.04	-3.61	0.95	4.40
Motivation	0.98	-1.97	1.03	-6.06	0.97	-0.53
Social Awareness	1.11	-8.26	1.16	-12.06	1.06	-3.74
Social Regulation	1.02	-1.87	1.01	-2.54	0.98	0.98

N.7 Spanish NA EQ Model Summary

The Spanish NA version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Spanish NA model.

O EQ Analysis for Swedish

This appendix presents a condensed version of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the Swedish version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 1092 respondents. Table O.1 presents the general statistics on means and standard deviations for the Swedish population, as well as for a Male/Female breakdown.

Table O.1: Swedish General Population Statistics
N = 1092

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	74.77	11.10	73.98	10.72	75.55	11.42
Self Regulation	70.73	12.67	72.67	11.69	68.80	13.31
Motivation	73.86	11.98	75.45	10.78	72.27	12.89
Social Awareness	75.15	10.95	71.65	10.92	78.464	9.82
Social Regulation	70.70	10.58	70.08	10.29	71.33	10.85

O.1 Swedish Inter-Item Correlation Analysis

This section presents the results of the inter-item correlation study for the Swedish version. The results are presented in Table O.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table O.2: Average Swedish Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.32	0.24	0.29	0.23	0.21
Item 2	0.32	0.25	0.27	0.22	0.23
Item 3	0.31	0.28	0.28	0.23	0.22
Item 4	0.32	0.26	0.29	0.25	0.22
Item 5	0.30	0.26	0.28	0.23	0.23
Item 6	0.31	0.25	0.30	0.25	0.21
Item 7	0.31	0.27	0.27	0.26	0.21
Item 8	0.30	0.27	0.26	0.23	0.21
Item 9	0.30	0.28	0.28	0.23	0.21
Item 10	0.31	0.27	0.26	0.26	0.21
Item 11		0.26	0.29	0.21	0.24
Item 12		0.27	0.27	0.23	
Mean	0.31	0.26	0.28	0.24	0.22
StDev	0.01	0.01	0.01	0.01	0.01

TTI SUCCESS INSIGHTS®

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the Swedish version based on inter-item correlation.

O.2 Swedish Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the Swedish version. The results are presented in Table O.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity based on internal structure.

Table O.3: Swedish Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.45	0.66	0.41	0.49	0.44
Item 2	0.41	0.57	0.57	0.55	0.33
Item 3	0.51	0.30	0.43	0.45	0.39
Item 4	0.42	0.52	0.35	0.33	0.35
Item 5	0.59	0.53	0.46	0.52	0.32
Item 6	0.49	0.60	0.26	0.33	0.44
Item 7	0.52	0.43	0.56	0.17	0.52
Item 8	0.54	0.36	0.62	0.47	0.45
Item 9	0.58	0.32	0.41	0.45	0.50
Item 10	0.46	0.35	0.62	0.20	0.45
Item 11		0.47	0.42	0.67	0.20
Item 12		0.41	0.56	0.53	
Mean	0.50	0.46	0.47	0.43	0.40
StDev	0.06	0.12	0.11	0.15	0.09

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table O.3, we have highlighted potential problem areas in red. These items are identified for monitoring and possible replacement at some point in the future. There is no alarm to be sounded at this time. The mean item-to-total correlation scores are are in an acceptable range and the standard deviations show consistent scoring.

O.3 Swedish Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the Swedish version. The results are presented in Table O.5. For convenience, Table 6.1 is reproduced as Table O.4.

Table O.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table O.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table O.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table O.5: Average Swedish α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.81	0.81	0.81	0.78	0.74
Lower CI	0.80	0.79	0.79	0.77	0.72
Upper CI	0.83	0.83	0.82	0.80	0.76

As is presented in Table O.5, all α values are in the Very Good ranges with the lone exception being the dip to 0.79 for the Social Regulation scale, implying a Respectable internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the Swedish version of EQ has solid evidence of internal consistency reliability.

Table O.6: α-if-item-deleted for Swedish

Est./Scale	SA	SR	MO	SoA	SoR
α	0.82	0.81	0.82	0.79	0.75
Item 1	0.81	0.78	0.82	0.77	0.73
Item 2	0.81	0.79	0.80	0.76	0.74
Item 3	0.80	0.81	0.81	0.77	0.74
Item 4	0.81	0.79	0.82	0.78	0.74
Item 5	0.79	0.79	0.81	0.76	0.75
Item 6	0.80	0.78	0.83	0.78	0.73
Item 7	0.80	0.80	0.80	0.80	0.72
Item 8	0.80	0.80	0.80	0.77	0.73
Item 9	0.79	0.81	0.81	0.77	0.72
Item 10	0.80	0.81	0.80	0.79	0.73
Item 11		0.80	0.82	0.75	0.76
Item 12		0.80	0.80	0.76	

O.4 Swedish Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

O.5 Swedish Factor Analysis

This section presents the results of the factor analysis study for the Swedish version. The results are presented in Tables O.7 through O.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table O.7: Swedish Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1			
Item 2		0.46	
Item 3		0.97	
Item 4	0.55		
Item 5	0.54		
Item 6			0.94
Item 7	0.54		
Item 8	0.59		
Item 9	0.58		
Item 10	0.35		0.41
CR	0.82		
VE	0.47		

Some comments on how to interpret the information presented in Tables O.7 through O.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor

TTI SUCCESS INSIGHTS®

pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table O.8: Swedish Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.40	0.56		
Item 2	0.76			
Item 3			0.48	
Item 4			0.78	
Item 5	0.59			
Item 6	0.36	0.37	0.40	
Item 7		0.39		0.43
Item 8		0.37		
Item 9		0.36		
Item 10				0.56
Item 11		0.44		
Item 12				0.60
CR	0.84			
VE	0.42			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table O.9: Swedish Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2		0.72		
Item 3		0.36		
Item 4				
Item 5			0.45	
Item 6	0.35			
Item 7	0.67			
Item 8	0.39		0.45	
Item 9	0.56			
Item 10		0.60		
Item 11			0.52	
Item 12				0.90
CR	0.85			
VE	0.42			

TTI SUCCESS INSIGHTS®

Table O.10: Swedish Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.62		
Item 2	0.40		0.66	
Item 3				
Item 4				
Item 5	0.40		0.64	
Item 6	0.36	0.41		
Item 7				
Item 8		0.82		
Item 9	0.43			
Item 10				0.96
Item 11	0.72			
Item 12	0.60			
CR	0.82			
VE	0.46			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table O.11: Swedish Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1		0.42		
Item 2				
Item 3			0.90	
Item 4		0.38		
Item 5		0.58		
Item 6		0.59		
Item 7	0.38			0.69
Item 8	0.36			
Item 9	0.73			
Item 10	0.52			
Item 11				
CR	0.76			
VE	0.40			

The main takeaway for the Swedish EQ model exploratory factor analysis is that the CR and VE values are all good or very close to good. There are no red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation



and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all EQ models at TTI Success Insights.

O.6 Swedish EQ Model Equating Functions

Table O.12: Swedish-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.94	6.92	0.98	2.01	0.96	4.17
Self Regulation	0.97	5.61	1.02	1.09	0.95	8.09
Motivation	0.93	8.19	1.03	0.31	0.86	13.42
Social Awareness	1.06	-7.12	1.09	-8.09	1.09	-10.10
Social Regulation	1.09	-6.03	1.14	-9.21	1.05	-2.99

Table O.13: Swedish-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.98	3.70	1.06	-3.45	0.90	9.91
Self Regulation	0.98	5.04	1.04	0.22	0.97	6.57
Motivation	0.98	4.55	1.09	-4.73	0.92	9.21
Social Awareness	1.11	-11.29	1.15	-12.51	1.15	-15.35
Social Regulation	1.15	-9.84	1.17	-10.84	1.10	-7.15

O.7 Swedish EQ Model Summary

The Swedish version of the EQ model shows strong evidence of internal consistency reliability and internal structure validity. While there are no red flags to be concerned about at this time, improvements to the questionnaire are forthcoming that should have a quite positive impact on the Swedish model.

P EQ Analysis for the United Kingdom

This appendix presents a condensed variation of the full analysis presented in Sections 10 through 11, synopsized in Section 15, for the English UK version of the TTI Success Insights EQ assessment. Due to a relatively small subset of data and the nearly identical language used in this assessment, one may consult the aforementioned sections to help answer any questions. The study is conducted on 1884 respondents. Table P.1 presents the general statistics on means and standard deviations for the English UK population, as well as for a Male/Female breakdown.

Table P.1: UK General Population Statistics
N = 1884

Scale	Gen Pop		M		F	
	Mean	StDev	Mean	StDev	Mean	StDev
Self Awareness	75.59	10.73	73.63	10.54	77.56	10.56
Self Regulation	67.49	12.43	69.21	11.95	65.76	12.68
Motivation	74.70	11.33	74.96	11.38	74.45	11.28
Social Awareness	72.98	11.98	69.48	11.94	76.47	10.95
Social Regulation	71.66	10.95	70.70	10.95	72.63	10.88

P.1 UK Inter Item Correlation Analysis

This section presents the results of the inter-item correlation study for the English UK version. The results are presented in Table P.2. As is discussed in Section 6, inter-item correlation is one possible measure of internal consistency reliability.

Table P.2: Average English-UK Inter-item Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.30	0.25	0.26	0.27	0.23
Item 2	0.29	0.26	0.24	0.25	0.25
Item 3	0.28	0.27	0.25	0.26	0.24
Item 4	0.28	0.25	0.26	0.27	0.25
Item 5	0.27	0.26	0.26	0.25	0.24
Item 6	0.29	0.26	0.27	0.27	0.24
Item 7	0.29	0.28	0.24	0.29	0.23
Item 8	0.27	0.26	0.25	0.26	0.22
Item 9	0.29	0.26	0.24	0.26	0.23
Item 10	0.29	0.28	0.24	0.27	0.23
Item 11		0.26	0.25	0.25	0.26
Item 12		0.27	0.25	0.25	
Mean	0.31	0.26	0.25	0.26	0.24
StDev	0.01	0.01	0.01	0.01	0.01

The results are quite solid overall. Recall that there are essentially two ranges for acceptable levels of inter-item correlation. The first is [0.20,0.40] and the second is [0.15,0.50]. Since all values are well within these limits, there are no issues identified for the English UK version based on inter-item correlation.

P.2 UK Corrected Item Total Correlation Analysis

This section presents the results of the item-total correlation study for the English UK version. The results are presented in Table P.3. As is discussed in Section 11, corrected item total correlation serves as a proxy for evidence of validity.

Table P.3: English-UK Corrected Item-total Correlations

Item/Scale	SA	SR	MO	SoA	SoR
Item 1	0.36	0.63	0.36	0.41	0.52
Item 2	0.44	0.48	0.50	0.55	0.28
Item 3	0.51	0.40	0.41	0.45	0.38
Item 4	0.49	0.56	0.34	0.42	0.34
Item 5	0.60	0.47	0.39	0.54	0.44
Item 6	0.41	0.50	0.28	0.40	0.41
Item 7	0.43	0.32	0.52	0.22	0.48
Item 8	0.58	0.48	0.48	0.43	0.56
Item 9	0.45	0.47	0.56	0.48	0.48
Item 10	0.42	0.32	0.56	0.39	0.47
Item 11		0.51	0.41	0.59	0.22
Item 12		0.36	0.47	0.59	
Average	0.47	0.46	0.44	0.46	0.42
StDev	0.07	0.09	0.09	0.10	0.09

With regard to item-to-total score correlation, there is little in the form of a specific range in the published literature that we have uncovered. The previous statement should be read with some caution. There are many acceptable ranges in the literature, but there is little in the form of consistency among them. Essentially, anything above 0.30 is considered good, with the mindset that the higher values the better the assessment items are.

However, there are limitations in the minds of these authors. If the correlations are too high, we may be finding ourselves in an overly homogeneous assessment. Generally speaking, we would like to see values in the 0.50 range or higher. We are not going to "throw" an item away for having a score below 0.50, but we would consider potential changes to assessment items in an attempt to raise the lower scores.

In Table P.3, we have highlighted potential problem areas in red. These four items are identified for monitoring and possible replacement at some point in the future. There is no alarm to be sounded at this time. The mean item-to-total correlation scores are are in an acceptable range and the standard deviations show consistent scoring.

P.3 UK Internal Consistency via the α Coefficient

This section presents the results of the internal consistency study for the English UK version. The results are presented in Table P.5. For convenience, Table 6.1 is reproduced as Table P.4.

Table P.4: Common Reliability Coefficient Interpretation, see [63]

Range	Interpretation
$\alpha \geq 0.90$	Consider Shortening the Scale
$0.80 \leq \alpha < 0.90$	Very Good
$0.70 \leq \alpha < 0.80$	Respectable
$0.65 \leq \alpha < 0.70$	Undesirable
$\alpha < 0.65$	Unacceptable

Note that there will be two different values of α for each scale. The first is presented in Table P.5 and is the measure of internal consistency presented with confidence intervals. The second is presented in Table C.6 and is intended to be used for comparison to the α-if-item-deleted statistic presented therein. Any differences in the two values are attributed to slightly different algorithmic approaches in the underlying statistical package used to compute them. The more stable the data, the less difference that is present.

Table P.5: Average UK α Coefficient Data

Est./Scale	SA	SR	MO	SoA	SoR
α	0.80	0.81	0.80	0.81	0.78
Lower CI	0.79	0.80	0.79	0.80	0.77
Upper CI	0.81	0.82	0.81	0.82	0.79

As is presented in Table P.5, all α values are in the Very Good ranges with the lone exception being the dip to 0.78 for the Social Regulation scale, implying a Respectable internal consistency. All these numbers are very solid and there is nothing to report based on the internal consistency values.

Regarding the α-if-item-deleted, a brief explanation is in order. This statistic is intended to measure the impact of each individual item on the internal consistency of the scale. The process is to remove the population responses to the individual item and the compute the α coefficient, hence the name α-if-item-deleted. What we are looking for is a drop or no movement in the α value when each item is removed. If the value of α goes up, the scale is potentially "better" without the item. If we find items for which this occurs, we flag the item for further analysis and monitoring.

In this case, there are no items identified as having an increase in the value of α at this time. We conclude that the English UK version of EQ has solid evidence of internal consistency reliability.

TTI SUCCESS INSIGHTS®

Table P.6: α-if-item-deleted for English UK

Est./Scale	SA	SR	MO	SoA	SoR
α	0.80	0.81	0.82	0.81	0.78
Item 1	0.79	0.78	0.79	0.80	0.75
Item 2	0.79	0.80	0.78	0.79	0.77
Item 3	0.78	0.80	0.80	0.80	0.76
Item 4	0.78	0.79	0.79	0.80	0.77
Item 5	0.77	0.80	0.80	0.79	0.76
Item 6	0.79	0.79	0.78	0.80	0.76
Item 7	0.79	0.81	0.78	0.82	0.75
Item 8	0.77	0.79	0.77	0.80	0.74
Item 9	0.78	0.80	0.77	0.79	0.75
Item 10	0.79	0.81	0.78	0.80	0.75
Item 11		0.79	0.79	0.78	0.78
Item 12		0.80	0.78	0.78	

P.4 UK Test-Retest Analysis

At the time of writing of this technical report, a sufficiently large sample of data was not available for a test-retest analysis. As soon as a relevant sample is obtained, this manual shall be updated with the test-retest analysis results.

P.5 UK Factor Analysis

This section presents the results of the factor analysis study for the English UK version. The results are presented in Tables P.7 through P.11. A complete discussion of the methods used is presented in Section 10, see Definition 10.1.

Table P.7: UK Factor Pattern Matrix
Self Awareness

Item/Factor	λ_1	λ_2	λ_3
Item 1			0.42
Item 2		0.56	
Item 3		0.88	
Item 4	0.57		
Item 5	0.54		
Item 6			0.81
Item 7	0.48		
Item 8	0.72		
Item 9	0.48		
Item 10			0.40
CR	0.80		
VE	0.44		

Some comments on how to interpret the information presented in Tables C.7 through C.11 is necessary. We are essentially looking for three main results. First, we wish to see a clean factor



pattern matrix, i.e., there is little or no cross-loading present. Cross-loading is a reference to each item having a nontrivial loading on more than one factor. Second, we wish to see balance in the factor pattern matrix, i.e., each factor has approximately the same number of items loading on it.

Table P.8: UK Factor Pattern Matrix
Self Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.60			
Item 2	0.55			
Item 3		0.50		
Item 4		0.77		
Item 5				0.93
Item 6	0.40			
Item 7			0.52	
Item 8			0.41	
Item 9	0.52			
Item 10			0.39	
Item 11	0.53			
Item 12			0.49	
CR	0.84			
VE	0.42			

Third, we wish to see reasonable levels of loading. Technically speaking, any factor load above, say, 0.35 is acceptable under certain conditions.

Table P.9: UK Factor Pattern Matrix
Motivation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1				
Item 2		0.82		
Item 3				
Item 4				
Item 5			0.94	
Item 6				
Item 7				0.92
Item 8	0.56			
Item 9	0.60			
Item 10		0.63		
Item 11	0.43			
Item 12	0.37			
CR	0.83			
VE	0.44			

TTI SUCCESS INSIGHTS®

Table P.10: UK Factor Pattern Matrix
Social Awareness

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1			0.57	
Item 2	0.75			
Item 3				
Item 4				0.95
Item 5	0.74			
Item 6		0.50		
Item 7	0.39			
Item 8			0.79	
Item 9	0.38	0.34		
Item 10	0.45			
Item 11		0.66		
Item 12	0.39	0.48		
CR	0.83			
VE	0.47			

The conditions may be summarized by considering the CR (Composite Reliability) and VE (Variance Explained) values at the bottom of the table. If $CR \geq 0.80$ and $VE \geq 0.50$, any factor pattern values above 0.35 are good.

Table P.11: UK Factor Pattern Matrix
Social Regulation

Item/Factor	λ_1	λ_2	λ_3	λ_4
Item 1	0.39			
Item 2				
Item 3			0.54	
Item 4	0.39			
Item 5	0.50			
Item 6	0.59			
Item 7		0.58		
Item 8	0.39		0.46	
Item 9		0.59		
Item 10		0.37		0.91
Item 11			0.37	
CR	0.80			
VE	0.39			

The main takeaway for the UK EQ model exploratory factor analysis is that the CR and VE values are mostly good or very close to good. There are some red flags. There is some lack of balance and some cross-loading that we have identified and we shall work, over time, to rectify this situation



and improve any VE values to get them above the magic 0.50 value. As is noted earlier in this technical manual, updates to the US English model are forthcoming and those changes are likely to have a very positive impact on all English based EQ models at TTI Success Insights.

P.6 UK EQ Model Equating Functions

Table P.12: U.K.-U.S. Equating Functions
Mean/StDev Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	0.97	3.56	1.00	2.01	0.96	4.17
Self Regulation	0.98	4.98	0.99	3.52	0.97	5.95
Motivation	0.98	3.42	0.97	4.86	0.99	2.29
Social Awareness	0.97	1.83	0.99	0.73	0.97	0.83
Social Regulation	1.06	-4.43	1.08	-5.04	1.04	-4.14

Table P.13: U.K.-U.S. Equating Functions
Equipercentile Method

Scale	Gen Pop		M		F	
	Slope	Intercept	Slope	Intercept	Slope	Intercept
Self Awareness	1.03	-1.46	1.06	-2.38	1.01	-0.26
Self Regulation	0.99	3.94	1.04	0.44	0.99	4.70
Motivation	1.04	-1.58	1.04	-0.22	1.05	-1.90
Social Awareness	1.02	-1.72	1.04	-2.80	1.04	-4.12
Social Regulation	1.09	-7.01	1.09	-6.07	1.10	-8.66

P.7 UK EQ Model Summary

The UK version of the EQ model shows signs of solid internal consistency reliability with some issues popping up in the internal structure validity. Updates to the US English based questionnaire show promise to have serious positive impact on all English versions of this model.

Q Assessment Adaptation Protocols

The International Test Commission's (ITC) stated goal is to assist in the exchange of information on test development and use among its members and affiliate organizations as well as with nonmember societies, organizations, and individuals who desire to improve test-related practices, see [131]. In support of this goal, the ITC has developed a series of guidelines ranging across a spectrum that includes comments on test use, quality control, test takers, and using tests for research, to name a few.

One of the major foci of the guidelines on test translation and adaptation, see [55], is to educate on the differences between translation and adaptation. The ITC states the following:

> Test translation is probably the more common term, but adaptation is the broader term and refers to moving a test from one language and culture to another. Test adaptation refers to all of the activities including: deciding whether or not a test in a second language and culture could measure the same construct in the first language; selecting translators; choosing a design for evaluating the work of test translators (e.g., forward and backward translations); ...

The guidelines are organized into six categories

1. Pre-conditions;
2. Test Development;
3. Confirmation [Empirical Analyses];
4. Administration;
5. Score Scales and Interpretation;
6. Documentation.

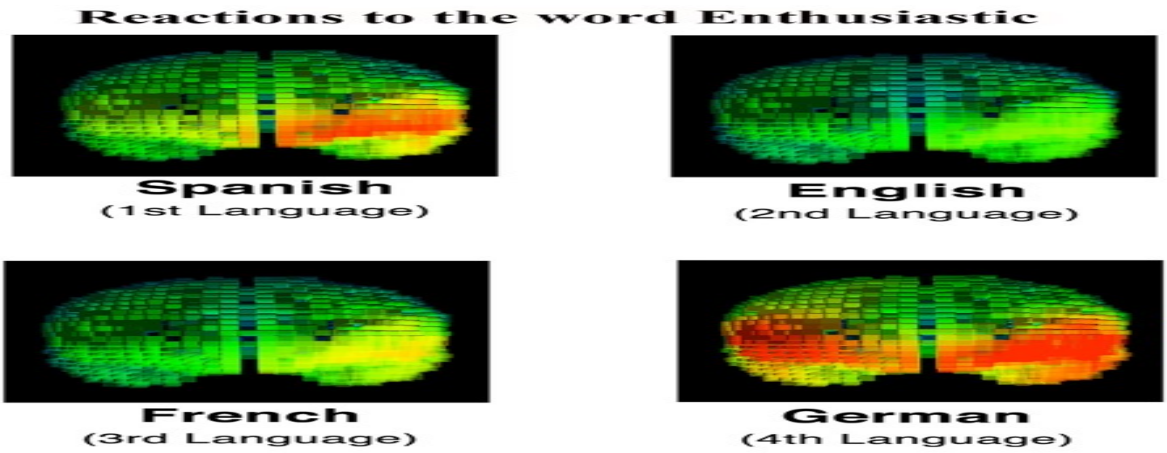


Figure Q.1: Emotional response to *enthusiastic*

Before going on to the guidelines, we present a short motivation for the importance of having validated test adaptation protocols. Figure Q.1 shows the emotional response of a multi-lingual

TTI SUCCESS INSIGHTS®

individual to the word *enthusiastic* presented to the person in multiple languages. This person's native language is Spanish (Mexico). The following explanation of the brain images is taken from [54].

> She showed the maximum positive response (left-side activation) to her native language (Spanish), a relatively neutral (no overt activation in either hemisphere) to English, a moderately positive response (some left side activation) to French, and a strong and complex response (activation on both sides) to German. This was consistent with the fact that she was more comfortable with languages like her own and found German difficult and requiring effort to understand.

Q.1 Pre-conditions

There are three pre-condition guidelines in [55], the first of which is legal in nature, and the second two are centered on ensuring what is being assessed is sufficiently understood in the same way across all cultural and language groups of interest.

PC-1 Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation.

Given that most of the suite of TTI Success Insights assessments are developed by TTI Success Insights, PC-1 should generally not be a problem. However, to cover all bases, it is necessary to at least document in some fashion that we do, in fact, have all necessary permissions.

PC-2 Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the population of interest is sufficient for the intended use (or uses) of the scores.

According to the explanation provided in [55], this guideline requires that what is assessed should be understood in the same way across language and cultural groups, and this is the foundation of valid cross-cultural comparisons. Since the ITC refers to this concept as being foundational, it is critical to measure and report this in an appropriate manner. The ITC offers suggestions that experts with respect to the construct measured and who are familiar with the cultural groups being tested should be recruited to evaluate the legitimacy of the construct measured in each of the cultural/linguistic groups.

PC-3 Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest.

The comments on this guideline from the ITC are centered around appropriate choices of translators. In summary, it is not enough to have individuals familiar with a language, even completely fluent in the target language. The translators must not only be native speakers of the target language, they must also be culturally knowledgeable, preferably native to the culture as well. It may be shown through the test adaptation process that it is not enough to have a single French translation, for example. It is necessary to determine whether large enough cultural differences exist between the various French speaking areas TTI Success Insights is associated with to have a significant impact on the interpretation or uses of the output of a given assessment. We may find that cultural differences dictate that the French language version of an assessment developed for France may not be suitable for French speaking Canadians. Regardless, the protocols should outline and discuss how to document whether cultural or linguistic differences exist and impact test adaptation.

Q.2 Test Development

TD-1 Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.

The ITC explanation for this guideline includes a discussion of how it is of major importance to have translators that are knowledgeable in areas of the two languages in question as well as cultural understanding, and even looking for those with subject matter expertise and general knowledge of test construction. There is also a discussion of using a minimum number of translators (at least 2) along with various designs (e.g., forward and backward translations).

TD-2 Use appropriate translation designs and procedures to maximize suitability of the test adaptation in the intended populations.

The main takeaway for this guideline is that the translation process should be focused on functionality of language, not literal translation. The language should feel as natural and acceptable in the target language as it does in the source language.

TD-3 Provide evidence that the test instructions and item content have similar meaning for all intended populations.

Note that this is the first guideline to use the term *evidence*. Our interpretation of that is a need for data collection and analysis in some form. The ITC suggests the following possibilities:

1. Use reviewers native to the local language and culture;
2. Use of samples of bilingual respondents;
3. Use of local surveys to evaluate the test;
4. Use of non-standard test administrations to increase acceptability and validity.

The ITC also suggests that using small tryouts, conducting interviews of the administrators and respondents, and cognitive interviewing have shown solid success in test adaptation protocol settings.

TD-4 Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.

The main concern expressed by the ITC in this setting is the possibility of confusion arising from the administration platform (e.g., an individual is not familiar with computerized testing) or rating scales (e.g., an individual has never seen a 5-point Likert scale) and similar for the other areas above. It is suggested that either practice problems are provided or a thorough briefing accompany the assessment in some form.

TD-5 Collect pilot data on the adapted test to enable item analysis, reliability assessment, and small-scale validity studies so that any necessary revisions to the adapted test can be made.

This guideline should be fairly well understood. We need to conduct validity and reliability studies on every language into which any of our assessments have been or will be translated. This is obviously a tall order and will take a good deal of time, especially for any languages which service smaller populations.

Q.3 Confirmation Guidelines

While there are only four main guidelines in the Confirmation section of [55], there is an incredible amount of information and insights covered under this heading.

C-1 Select samples with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses.

This appears to be a bit of an obvious statement. However, the subtleties underlying the statement, and their implications, may or may not be as obvious. As noted in the first part of this series, there are some fairly strict guidelines on how to translate or adapt a test from one language to another. It would seem that if we require the target language translator to be native to the language and the culture, then the test samples should also be native to the language and the culture. In other words, if the target language is Spanish from the Americas, it would not be acceptable to take samples from Spain or to find a Spanish speaking population in the U.S.

[55] also makes some interesting comments regarding reliability studies. We provide the following quote from pp. 17:

> However, it is often overlooked that reliability is a joint characteristic of the test and the population (McDonald, 1999) - because it depends on both the true score variance (population characteristic) and error variance (test characteristic). Therefore, the same error variance can lead to a higher reliability simply due to the larger true score variance in the source language group. (McDonald, 1999) shows that the Standard Error of measurement (which is the square root of error variance) is in fact a more appropriate quantity to compare between samples, not reliability.

It is noted that the reference to (McDonald, 1999) is not provided in the attached bibliography. At the time of writing this article, the source [55] unfortunately does not provide the full reference in its bibliography. The authors of this article are researching to find the exact source.

The Guidelines provide the following rough guidance on minimal sample sizes:

1. Identification of potentially biased test items require a minimum of 200 persons per version of the test;
2. Item Response Theory (IRT) analyses require a minimum of 500 persons per version of the test;
3. Factor analytic studies require a minimum of 300 persons per version of the test.

The authors of this article do note that in many studies conducted by the TTI research team have shown that samples of 1000 still show some signs of instabilities in the convergence of IRT models in Likert-style response formats. As such, we typically require a minimum of 1200 persons per version of the test. It is also noted that in some cases this is a very difficult level of participation to achieve and limited information may be available for smaller sample sizes.

C-2 Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.

Before discussing the area of item equivalence, note that the approaches for construct and method equivalence are discussed in part one of this series of articles under guidelines **PC-2** and **PC-3**, respectively.

Item equivalence is essentially studied under the name *differential item functioning (DIF) analysis*. According to [55],

> In general, DIF exists if two test-takers, from two different (cultural-linguistic) populations, have the same level of the measured trait but have different response probability on a test item.

From the perspective of the TTI research team, one can replace the statement "different (cultural-linguistic) populations" with "categories of populations" in the sense that in the U.S. we have a need to study the existence of *disparate impact* assessment items may have on protected classes (categories) such as race, gender, age, and so on.

[55] points to two specific areas as potential sources of DIF: translation problems and cultural differences. As an example of a translation problem, it is necessary to avoid the use of vocabulary in the target language that is correct in meaning but not commonly used in a subset of the target language local culture. The following is a partial list of potential problem areas:

1. Change in item difficulty due to sentence length, complexity, or use of easy or difficult vocabulary;
2. Change of meaning due to deletion of sentences, inaccurate translation, more than one meaning in target language;
3. Non-equivalent meaning of words across cultures;
4. Cultural differences causing items to function different across languages.

[55] provides the following guidance on procedures to evaluate DIF.

1. IRT based methods;
2. Mantel-Haenszel (MH) procedure and extensions;
3. Logistic regression (LR) based procedures;
4. Restricted factor analytic (RFA) procedures.

While it was noted earlier that construct equivalence is covered, in part, in **PC-2**, [55] states the following additional information about construct equivalence:

> Finally, yet importantly, this guideline will require researchers to address construct equivalence. There are at least four statistical approaches for assessing construct equivalence across source and target language versions of a test: Exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and comparison of nomological networks (Sireci, Patsula, & Hambleton, 2005).

The reference for (Sireci, Patsula, & Hambleton, 2005) is provided in [162]. It is noted in [55] that there are no commonly agreed-upon rules for deciding when structures can be considered equivalent. Hence, as EFA does not accommodate separate factor structures, it is more desirable to use CFA, weighted MDS (WMDS), or Exploratory Structural Equation Modeling (ESEM).

C-3 Provide evidence supporting the norms, reliability, and validity of the adapted test version of the test in the intended populations.

Directly from [55]:

> The norms, validity evidence, and reliability evidence of a test in its source language version do not automatically apply to other possible adaptations of the test in to different cultures and languages.

Further, [55] suggests essentially following the standards set forth in [4] as well as in [163]. These standards state that one should provide evidence of validity based on the following five areas:

1. Test content;
2. Internal structure;
3. Relations to other variables;
4. Response processing;
5. Consequences of testing.

[55] also provides a short list of approaches to internal structure validity evidence such as EFA, CFA, structural equation modeling (SEM), and multitrait-multimethod analyses.

C-4 Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.

This topic may be more readily understood in the language of parallel form tests. For example, given a high stakes testing environment, there may be an increased risk of cheating. In such cases, multiple versions of the test may be produced to lessen the probability of an individual successfully cheating. However, how does one know if a given score on, say, Form A is equivalent to that same score on Form B? It may be the case that one form or the other is more or less difficult than its counterpart(s). As an example, it may be the case that a score of 70% on Form A is equivalent to a score of 60% on Form B due to an increased level of difficulty on Form B.

The process of determining the equivalence of scores in this example is called equating scores. If we are looking at different adaptations of a test into other languages and/or cultures, the process is called linking the scores. In order to establish the linking process, [55] suggests it is best to address the following three questions:

1. Is there evidence that the same construct is being measured in the source and target language versions of the test? Does the construct have the same relationship with other external variables in the new culture?
2. Is there strong evidence that sources of method bias have been eliminated?
3. Is the test free of potentially biased test items?

Q.4 Administration Guidelines

While there are only two main guidelines in the Administration section of [55], there is an incredible amount of information and insights covered under this heading.

A-1 Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect validity of the inferences drawn from the scores.

The idea here is to implement the administration guidelines in such a manner as to address all factors that may threaten the validity of test scores in specific cultural and linguistic contexts. Examples include being aware of any aspects of instructions that may have posed problems in the original language or in other adapted versions. The key to successfully addressing this guideline is being prepared to anticipate potential problems and posing potential solutions a priori.

Some areas of concern may be clarity of instructions, response mechanisms, allowable time, motivation to complete, knowledge of the purpose, and how the assessment is to be scored.

A-2 Specify testing conditions that should be followed closely in all populations of interest.

This guideline is all about establishing standard procedures that all assessment respondents should be required to adhere to. Some accommodations may be warranted in specific circumstances in order to ensure that each respondent has the best opportunity to respond to the assessment at their best, whatever that may mean.

Note that any variations should be noted so that these differences may be accounted for and analyzed as necessary. Their impact on generalizations and interpretations may or may not be significant. Some important aspects of the assessment process include ensuring testing instructions and related procedures are adapted in a standardized way which is suitable to the new language and culture and ensuring that administrators are trained on the new procedures. It is also highly important for the administrators to **respect the new procedures and not to "hold on" to the old guard.**

Q.5 Score Scales and Interpretation Guidelines

While there are only two main guidelines in the Score Scales and Interpretation section of [55], there is an incredible amount of information and insights covered under this heading.

SSI-1 Interpret any group score differences with reference to all relevant available information.

While this guideline is part of the ITC adaptation guidance, this advice is pertinent and relevant to anyone in the assessment industry. One should not take score differences across groups to be a one-off comparison. All relevant possibilities for differences in scoring should be taken into account regardless of whether there are differences in language and culture or not.

This is a particularly relevant point in the U.S. given the overly litigious society in existence here. In multiple U.S. Supreme Court Decisions, the Court has held that differences in assessment scale scores alone is not enough to establish a disparate impact claim, although many plaintiffs over the years have attempted to claim that these differences should be sufficient for such claims. We do not present specific cases, although the interested reader may contact the authors for a list of such cases and decisions.

Some prime examples of reasons that score differences may exist, and that need to be ruled out prior to drawing conclusions, include differential motivation, less effective education systems, lack of familiarity with testing/assessment procedures, and understanding the purpose of the assessment.

SSI-2 Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.

This section is at least partially linked to an earlier topic called "linking" or "equating". This is meant to establish evidence that construct, method, and item bias are not present in the adapted version of the assessment. According to [55], pp. 27,

> ...valid direct score comparisons can only be done when scores show the highest level of equivalence, called scalar equivalence or full score equivalence, which requires scales in each group have the same measurement unit and the same origin across groups.

The authors of [55] state that the key point is to not over-interpret assessment scores. For example, do not make comparative statements about respondent performance across language groups unless measurement invariance has been established. ß

Q.6 Documentation Guidelines

Finally, there are only two main guidelines in the Documentation section of [55]. However, there is once again an incredible amount of information and insights covered under this heading.

Doc-1 Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population.

Documentation should contain enough detail that future researchers are able to replicate the procedures to confirm findings or for use on other populations. These documents should include all quantitative and qualitative evidence associated with the adaptation process.

Doc-2 Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population.

This documentation should be written for the test user to be used in a practical setting, consistent with the ITC Guidelines on Test Use. For example, such a user manual should contain, at a minimum, the following:

1. Description of the construct(s) measured and summarize the information, include a description of the adaptation process;
2. Summaries of evidence supporting the adaptation, including cultural suitability of content, instructions, response format, etc.;
3. Definitions of suitability of use of test within various subgroups within the population and any restrictions on use;
4. Explanations of issues that may need to be considered in relation to good practice and test administration;
5. Explanation of whether and how inter-population comparisons may be made;
6. Information necessary for scoring and norming, or a description of how users may access scoring procedures;
7. Guidelines for the interpretation of results, including information on the implications of validity and reliability of the data on the inferences that may be drawn from the scores.

R Assessment Development Outline

