TTI Success Insights
Style Insights® Technical Manual Version 1.0

Eric Gehrig, PhD*
Vice President
Research & Development
Target Training International, Ltd.

Ron Bonnstetter, PhD†
Senior Vice President
Research & Development
Target Training International, Ltd.

February 23, 2021

*PhD, Mathematics, Arizona State University, 2007.
†Professor Emeritus, University of Nebraska-Lincoln.

TTI SUCCESS INSIGHTS®

Table of Contents

List of Figures vi

List of Tables viii

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 . . . 3
- 1.4 Analyses of the TTI Success Insights Style Insights Assessment for Diagnostic Purposes 5

2 The History and Theory of DISC 7
- 2.1 Pre-Prescot Lecky . . . 7
- 2.2 Walter Vernon Clarke . . . 16
- 2.3 Understanding the Assessment Creation Process . . . 17
- 2.4 John Cleaver: From AVA to DISC . . . 18
- 2.5 John G. Geier's Contribution . . . 18
- 2.6 The Creation and Expansion of Target Training International, Inc. . . . 19
- 2.7 Returning to the Issue of Average Graph Values . . . 20
- 2.8 History of DISC Concluding Comments . . . 23

3 A Few Comments on Reliability and Validity 24
- 3.1 Thoughts on Evidence of Reliability . . . 24
- 3.2 Thoughts on Evidence of Validity . . . 27
- 3.3 Correlation v. Causation . . . 33
- 3.4 Understanding Self-report Assessment Bias . . . 34
- 3.5 Forced-rank v. Likert-style Assessment Formats . . . 36
- 3.6 Parting Thoughts . . . 39

4 Item Difficulty and Discrimination 41
- 4.1 A Short Review of the American Psychological Association's View of Item Analysis . 41
- 4.2 TTI Success Insights Style Insights Forced-rank Item Analysis Results . . . 42
- 4.3 Item Analysis Summary . . . 44

5 Average Inter-item Correlation 45
- 5.1 A Brief Description of Average Inter-Item Correlation . . . 45
- 5.2 Inter Item Correlation Results . . . 46
- 5.3 The Attenuation Paradox . . . 47

6 Corrected Item Total Correlation 49
- 6.1 A Brief Discussion on Corrected Item Total Correlation . . . 49
- 6.2 Corrected Item-Total Correlation Results . . . 50
- 6.3 Corrected Item Total Correlation Summary . . . 51

7 Internal Consistency as a Measure of Reliability 53
- 7.1 A Short Review of the American Psychological Association's View of Reliability . . . 54
- 7.2 A Brief Discussion of Coefficient α . . . 54
- 7.3 A Brief Discussion of Coefficient ω . . . 56

7.4 α-if-Item-Deleted Analysis . . . 57
7.5 A Note on Comparing Reliability and Validity . . . 58

8 Temporal Consistency as a Measure of Reliability 60
8.1 A Brief Review of the APA's View of Temporal Consistency . . . 60
8.2 TTI Success Insights Test-Retest Studies . . . 61
8.3 TTI Success Insights Test-Retest Summary . . . 63

9 Generalizability Theory Applied to Test-retest Data 64
9.1 The Basics of One- and Two-Facet Generalizability Theory Models . . . 64
9.2 One-Facet G-theory Model Applied to TTI Success Insights Style Insights Test-Retest Data . . . 69
9.3 G-theory Summary . . . 72

10 Exploratory Factor Analysis 74
10.1 A Brief History of Factor Analysis . . . 74
10.2 Geometric Interpretation of Correlation . . . 75
10.3 The Basic Factor Model . . . 81
10.4 Factor Analysis in Matrix Form . . . 84
10.5 A Condensed View of the MinRes Problem . . . 84
10.6 Factor Scores and Indeterminacy . . . 88

11 A Graded Response Theory Approach to Likert Response Format DISC 90
11.1 From Classical Test Theory to Item Response Theory . . . 90
11.2 A Short Overview of the Rasch Models . . . 93
11.3 The Transition from Dichotomous to Free Response Models . . . 98
11.4 Item and Test Information . . . 102
11.5 Parameter Estimation and Scoring Rubric . . . 107

12 Thurstonian Item Response Theory Approach to Forced-rank Assessments 109
12.1 The Basics of Thurstonian Item Response Theory Models . . . 112
12.2 Item Characteristic Functions . . . 117
12.3 Item and Test Information Functions . . . 118

13 Relationships to Other Variables Evidence of Validity 120
13.1 Talent Insights Data and the Appropriate Subsets . . . 120
13.2 Internal Consistency Estimates for Talent Insights Data Subsets . . . 123
13.3 A Brief Review of Logistic Regression . . . 127
13.4 Confusion Matrices, ROC, and AUC . . . 129
13.5 A Brief Introduction to Principal Component Analysis . . . 132
13.6 Job Description Classification Results . . . 133
13.7 Summary and Future Work . . . 137
13.8 Identification Results . . . 138
13.9 Additional Results . . . 143

14 Response Processing Evidence of Validity 144
14.1 Using Electroencephalograph (EEG) to Measure Response Processing . . . 144
14.2 Response Processing Summary . . . 147

15 Consequences of Testing Evidence of Validity **148**
15.1 The Intended Purpose of TTI Success Insights' Style Insights Assessment 148
15.2 Addressing Consequences of Testing 148
15.3 Consequences of Testing Summary 149

16 Summary and Future Work **150**
16.1 Internal Consistency Reliability 150
16.2 Temporal Consistency Reliability 150
16.3 Alternate Forms Reliability 151
16.4 Generalizability Theory 151
16.5 Test Content Validity Evidence 151
16.6 Internal Structure Validity Evidence 151
16.7 Relationships to External Variables Validity Evidence 152
16.8 Response Processing Validity Evidence 152
16.9 Consequences of Testing Validity Evidence 153
16.10 TTI Success Insights Continual Improvement Process 154
16.11 Overall Summary and Future Work 154

References **157**

A Assessment Adaptation Protocols **168**
A.1 Pre-conditions 169
A.2 Test Development 170
A.3 Confirmation Guidelines 171
A.4 Administration Guidelines 173
A.5 Score Scales and Interpretation Guidelines 174
A.6 Documentation Guidelines 175

B Relationships to External Variables Results **176**
B.1 Demographic Breakdown of the Data 176
B.2 Model Results: No Holdout Samples 180
B.3 Model Results with Holdout Samples 206

C Historical Item Difficulty and Discrimination Charts **221**
C.1 2012 Historical Item Analysis 221
C.2 2013 Historical Item Analysis 223
C.3 2017 Historical Item Analysis 225

D Historical Inter-item Correlation Charts **227**
D.1 2012 Historical Inter-item Correlation 227
D.2 2013 Historical Inter-item Correlation 228
D.3 2017 Historical Inter-item Correlation 229

E Historical Item-Total CorrelationCharts **230**
E.1 2012 Historical Corrected Item-total Correlation 230
E.2 2013 Historical Corrected Item-total Correlation 231
E.3 2017 Historical Corrected Item-total Correlation 232

F Internal Consistency Tables **233**
F.1 2012 Historical Internal Consistency 233
F.2 2013 Historical Internal Consistency 233
F.3 2017 Historical Internal Consistency 233

G α-if-item-deleted Tables **234**
G.1 2012 Historical α-if-item-deleted 234
G.2 2013 Historical α-if-item-deleted 236
G.3 2017 Historical α-if-item-deleted 238

H Assessment Development Outline **240**

List of Figures

2.1 Average, Adapted, & Natural Graph Comparison . . . 20
2.2 Neurological Support for Two Graphs . . . 22
2.3 S-Loreta Image of Gamma Activity in the Frontal Lobes . . . 22
2.4 TTI Success Insights Experiment . . . 22
3.1 Bad Correlation v Causation Example . . . 33
4.1 Information v. Target Trait . . . 42
4.2 Forced-rank Dominance . . . 43
4.3 Forced-rank Influence . . . 43
4.4 Forced-rank Steadiness . . . 43
4.5 Forced-rank Compliance . . . 44
5.1 English US Style Insights Inter-item Correlation Plot . . . 46
6.1 English US Style Insights Corrected Item-total Correlation Plot . . . 50
10.1 Vectors in $\mathbb{R}^n$ . . . 76
10.2 Geometric Interpretation of Factor Analysis Parameters . . . 88
11.1 Generic CDF . . . 93
11.2 Example of Item Difficulty . . . 94
11.3 Example of Item Discrimination . . . 95
11.4 Rasch 1-PL Model . . . 96
11.5 Rasch 2-PL Model . . . 97
11.6 Rasch 3-PL Model . . . 98
11.7 Example Boundary Characteristic Curves . . . 103
11.8 Example Item Response Category Characteristic Curves . . . 104
11.9 Example Item Information Function . . . 105
11.10 Example Test Information Function . . . 106
11.11 Example Item Information Function with Reliability Boundary . . . 107
11.12 Example Test Information Function with Reliability Boundary . . . 107
12.1 Generic Item Characteristic Surface for Thurstonian IRT Model . . . 118
12.2 Generic Item Information Surface for Thurstonian IRT Model . . . 119
13.1 Confusion Matrix . . . 130
13.2 Example ROC Curve . . . 131
13.3 Natural Dominance Log Odds Plot . . . 134
13.4 Intentional Log Odds Plot . . . 134
13.5 CEO ROC Curve: All Variables . . . 136
13.6 Sales1 ROC Curve: All Variables . . . 139
13.7 Accountant ROC Curve: All Variables . . . 140
13.8 Manager ROC Curve: All Variables . . . 141
13.9 Sales2 ROC Curve: All Variables . . . 142
14.1 An Example of Asymmetry in EEG Captured Brain Activity . . . 145
14.2 Confirmation Between the Survey and Brain Response Respondent Answer: 4 . . . 146
14.3 Possible Socially Acceptable Response Based on Brain Activity Respondent Answer: 5 . . . 146
14.4 Brain Response to Confusing Items Respondent Answer: 3 . . . 146
14.5 Brain Response to Double Negative Items Respondent Answer: 5 . . . 147

14.6 Neutral Brain Reaction Respondent Answer: 5 . . . 147
A.1 Emotional response to *enthusiastic* . . . 168
C.1 2012 Forced-rank Dominance . . . 221
C.2 2012 Forced-rank Influence . . . 221
C.3 2012 Forced-rank Steadiness . . . 222
C.4 2012 Forced-rank Compliance . . . 222
C.5 2013 Forced-rank Dominance . . . 223
C.6 2013 Forced-rank Influence . . . 223
C.7 2013 Forced-rank Steadiness . . . 224
C.8 2013 Forced-rank Compliance . . . 224
C.9 2017 Forced-rank Dominance . . . 225
C.10 2017 Forced-rank Influence . . . 225
C.11 2017 Forced-rank Steadiness . . . 226
C.12 2017 Forced-rank Compliance . . . 226

List of Tables

1.1 TTI Success Insights EQ Technical Manual Navigation . . . 2
1.2 TTI Success Insights Style Insights Scale Abbreviation Table . . . 3
3.1 Common Reliability Coefficient Interpretation . . . 26
4.1 TTI Success Insights Information v. Trait Example . . . 42
5.1 Average English US Inter-item Correlations TTI SI Style Insights . . . 47
6.1 Corrected Item-total Correlation Levels Suggested by APA . . . 49
6.2 English US Corrected Item-total Correlations TTI SI Style Insights . . . 51
7.1 Average English US α Coefficient Data TTI SI Style Insights . . . 55
7.2 Common Reliability Coefficient Interpretation, see [59] . . . 55
7.3 Average English US ω Coefficient Data TTI SI Emotional Quotient . . . 57
7.4 English US α-if-Item-Deleted Data Dominance, $\alpha = 0.889$ . . . 57
7.5 English US α-if-Item-Deleted Data Influence, $\alpha = 0.865$ . . . 58
7.6 English US α-if-Item-Deleted Data Steadiness, $\alpha = 0.855$ . . . 58
7.7 English US α-if-Item-Deleted Data Compliance, $\alpha = 0.849$ . . . 58
8.1 Test-Retest Data Breakdown TTI Success Insights Style Insights . . . 62
8.2 Test-Retest Correlations 01/01/2015 to 03/25/2020 TTI Success Insights Style Insights . . . 62
9.1 Hypothetical Example of 5 Individuals and a Single Facet . . . 66
9.2 Schematic View of Teacher Lesson Plan Example . . . 67
9.3 Style Insights Scale X Test-Retest as a Two-Facet G-Theory Model . . . 69
9.4 Hypothetical Example of n Individuals and One Facet ("Time") . . . 70
9.5 Style Insights Dominance Scale Test-Retest One-Facet G-Theory Model 1/1/2015 to 3/25/2020 . . . 70
9.6 Style Insights Influence Scale Test-Retest One-Facet G-Theory Model 1/1/2015 to 3/25/2020 . . . 71
9.7 Style Insights Steadiness Scale Test-Retest One Facet G-Theory Model 1/1/2015 to 3/25/2020 . . . 71
9.8 Style Insights Compliance Scale Test-Retest One-Facet G-Theory Model 1/1/2015 to 3/25/2020 . . . 72
9.9 Style Insights One-Facet Test-Retest Generalizability and Dependability 1/1/2015 to 3/25/2020 . . . 72
10.1 Examples of Observed Variables from [100] . . . 81

10.2 Composition of Variance for Full Common Factor Model . . . 83
11.1 Response Frequencies Dichotomization Example . . . 91
11.2 Response Distribution Dichotomization Example . . . 92
13.1 Most Commonly Chosen Job Descriptions: Talent Insights . . . 121
13.2 Naming Convention . . . 122
13.3 Basic Information Sales1 . . . 122
13.4 Basic Information Accountants . . . 122
13.5 Basic Information Manager . . . 123
13.6 Basic Information CEO . . . 123
13.7 Basic Information Sales2 . . . 123
13.8 Common Reliability Coefficient Interpretation, see [59] . . . 124
13.9 Style Insights Statistics: Sales1 . . . 124
13.10 Motivation Insights Statistics: Sales1 . . . 124
13.11 Style Insights Statistics: Accountant . . . 125
13.12 Motivation Insights Statistics: Accountant . . . 125
13.13 Style Insights Statistics: Managers . . . 125
13.14 Motivation Insights Statistics: Managers . . . 126
13.15 Style Insights Statistics: Sales2 . . . 126
13.16 Motivation Insights Statistics: Sales2 . . . 126
13.17 Style Insights Statistics: CEO . . . 126
13.18 Motivation Insights Statistics: CEO . . . 127
13.19 Example Confusion Matrix . . . 131
13.20 CEO Sample Area Under Curve . . . 135
13.21 CEO Sample Brier Statistic . . . 135
13.22 CEO v Entrepreneur Comparison . . . 136
13.23 Sales1 Sample Area Under Curve . . . 138
13.24 Sales1 Sample Brier Statistic . . . 139
13.25 Accountant Sample Area Under Curve . . . 139
13.26 Accountant Sample Brier Statistic . . . 140
13.27 Manager Sample Area Under Curve . . . 141
13.28 Manager Sample Brier Statistic . . . 141
13.29 Sales2 Sample Area Under Curve . . . 142
13.30 Sales2 Sample Brier Statistic . . . 142

B.1 TTI Success Insights Demographics
Gender, N = 549,782 . . . 177
B.2 TTI Success Insights Demographics
Ethnicity, N = 155,892 . . . 177
B.3 TTI Success Insights Demographics
Decade Born, N = 125,750 . . . 177
B.4 TTI Success Insights Demographics
Education (US), N = 155,758 . . . 178
B.5 TTI Success Insights Demographics
Employment Status, N = 157,419 . . . 178
B.6 TTI Success Insights Demographics
Disability Status, N = 155,472 . . . 178
B.7 TTI Success Insights Demographics
Veteran Status, N = 155,472 . . . 179
B.8 Acute Care Nurses Sample
Area Under Curve, N = 136 . . . 180
B.9 Advertising & Promotions Managers Sample
Area Under Curve, N = 220 . . . 180
B.10 Agricultural Workers, All Other Sample
Area Under Curve, N = 320 . . . 180
B.11 Airline Pilots, Copilots, & Flight Engineers Sample
Area Under Curve, N = 212 . . . 180
B.12 Architectural & Engineering Managers Sample
Area Under Curve, N = 167 . . . 180
B.13 Art Directors Sample
Area Under Curve, N = 201 . . . 181
B.14 Assemblers & Fabricators, All Other Sample
Area Under Curve, N = 411 . . . 181
B.15 Auditors Sample
Area Under Curve, N = 148 . . . 181
B.16 Automotive Body & Related Repairers Sample
Area Under Curve, N = 175 . . . 181
B.17 Automotive Engineers Sample
Area Under Curve, N = 128 . . . 181
B.18 Automotive Service Technicians & Mechanics Sample
Area Under Curve, N = 120 . . . 182
B.19 Baristas Sample
Area Under Curve, N = 109 . . . 182
B.20 Bartenders Sample
Area Under Curve, N = 110 . . . 182
B.21 Bill & Account Collectors Sample
Area Under Curve, N = 161 . . . 182
B.22 Billing & Posting Clerks Sample
Area Under Curve, N = 126 . . . 182
B.23 Budget Analysts Sample
Area Under Curve, N = 125 . . . 183
B.24 Building Cleaning Workers, All Other Sample
Area Under Curve, N = 112 . . . 183

B.25 Business Intelligence Analysts Sample
Area Under Curve, N = 287 . . . 183
B.26 Carpenters Sample
Area Under Curve, N = 368 . . . 183
B.27 Cashiers Sample
Area Under Curve, N = 155 . . . 183
B.28 Chemists Sample
Area Under Curve, N = 126 . . . 183
B.29 Child, Family, & School Social Workers Sample
Area Under Curve, N = 247 . . . 184
B.30 Chiropractors Sample
Area Under Curve, N = 317 . . . 184
B.31 Clergy Sample
Area Under Curve, N = 317 . . . 184
B.32 Combined Food Preparation &
Serving Workers, Including Fast Food Sample
Area Under Curve, N = 317 . . . 184
B.33 Compliance Managers Sample
Area Under Curve, N = 215 . . . 184
B.34 Computer & Information Research Scientists Sample
Area Under Curve, N = 134 . . . 185
B.35 Computer Network Support Specialists Sample
Area Under Curve, N = 306 . . . 185
B.36 Computer Programmers Sample
Area Under Curve, N = 487 . . . 185
B.37 Computer Systems Analysts Sample
Area Under Curve, N = 360 . . . 185
B.38 Construction Laborers
Area Under Curve, N = 114 . . . 185
B.39 Cooks, All Other
Area Under Curve, N = 121 . . . 185
B.40 Counselors, All Other
Area Under Curve, N = 240 . . . 186
B.41 Counter & Rental Clerks
Area Under Curve, N = 127 . . . 186
B.42 Credit Analysts
Area Under Curve, N = 127 . . . 186
B.43 Demonstrators & Product Planners
Area Under Curve, N = 477 . . . 186
B.44 Dentists, All Other Specialists
Area Under Curve, N = 146 . . . 186
B.45 Dentists, General
Area Under Curve, N = 386 . . . 186
B.46 Designers, All Other
Area Under Curve, N = 433 . . . 187
B.47 Directors, Religious Activities & Education
Area Under Curve, N = 199 . . . 187

B.48 Door-to-Door Sales Workers, News & Street Vendors, & Related Workers
Area Under Curve, N = 199 . . . 187
B.49 Driver-Sales Workers
Area Under Curve, N = 169 . . . 187
B.50 Electrical Engineers
Area Under Curve, N = 169 . . . 187
B.51 Elementary School Teachers, Except Special Education
Area Under Curve, N = 363 . . . 188
B.52 Emergency Medical Technician & Paramedics
Area Under Curve, N = 139 . . . 188
B.53 Engineers, All Other
Area Under Curve, N = 354 . . . 188
B.54 Entertainers and Performers, Sports & Related Workers, All Other
Area Under Curve, N = 104 . . . 188
B.55 Environmental Engineers
Area Under Curve, N = 134 . . . 188
B.56 Environmental Scientists & Specialists, Including Health
Area Under Curve, N = 110 . . . 189
B.57 Family & General Practitioners
Area Under Curve, N = 176 . . . 189
B.58 Financial Clerks, All Other
Area Under Curve, N = 110 . . . 189
B.59 Financial Managers
Area Under Curve, N = 484 . . . 189
B.60 First-Line Supervisors of Construction Trades & Extraction Workers
Area Under Curve, N = 279 . . . 189
B.61 First-Line Supervisors of Non-Retail Sales Workers
Area Under Curve, N = 338 . . . 190
B.62 First-Line Supervisors of Office & Administrative Support Workers
Area Under Curve, N = 191 . . . 190
B.63 First-Line Supervisors of Police & Detectives
Area Under Curve, N = 130 . . . 190
B.64 First-Line Supervisors of Production & Operating Workers
Area Under Curve, N = 208 . . . 190
B.65 First-Line Supervisors of Transportation & Material-Moving Machine & Vehicle Operators
Area Under Curve, N = 171 . . . 190
B.66 Food Preparation & Serving Related Workers, All Other
Area Under Curve, N = 140 . . . 191
B.67 Food Processing Workers, All Other
Area Under Curve, N = 144 . . . 191
B.68 Fundraisers
Area Under Curve, N = 172 . . . 191
B.69 Graphic Designers
Area Under Curve, N = 449 . . . 191

B.70 Ground Maintenance Workers, All Other
Area Under Curve, N = 138 . . . 191
B.71 Hairdressers, Hairstylists, & Cosmetologists
Area Under Curve, N = 115 . . . 192
B.72 Health Technologists & Technicians, All Other
Area Under Curve, N = 122 . . . 192
B.73 Healthcare Practitioners & Technical Workers, All Other
Area Under Curve, N = 327 . . . 192
B.74 Heating, Air Conditioning, & Refrigeration Mechanics & Installers
Area Under Curve, N = 107 . . . 192
B.75 Industrial Production Managers
Area Under Curve, N = 196 . . . 192
B.76 Information Security Analysts
Area Under Curve, N = 153 . . . 193
B.77 Installation, Maintenance, & Repair Workers, All Other
Area Under Curve, N = 498 . . . 193
B.78 Insurance Policy Processing Clerks
Area Under Curve, N = 105 . . . 193
B.79 Insurance Underwriters
Area Under Curve, N = 477 . . . 193
B.80 Interior Designers
Area Under Curve, N = 242 . . . 193
B.81 Investment Underwriters
Area Under Curve, N = 126 . . . 194
B.82 Legal Support Workers, All Other
Area Under Curve, N = 365 . . . 194
B.83 Librarians
Area Under Curve, N = 123 . . . 194
B.84 Licensed Practical & Licensed Vocational Nurses
Area Under Curve, N = 125 . . . 194
B.85 Life Scientists, All Other
Area Under Curve, N = 114 . . . 194
B.86 Life, Physical, & Social Science Technicians, All Other
Area Under Curve, N = 146 . . . 195
B.87 Logistics Managers
Area Under Curve, N = 241 . . . 195
B.88 Maintenance & Repair Workers, General
Area Under Curve, N = 261 . . . 195
B.89 Manufacturing Engineers
Area Under Curve, N = 231 . . . 195
B.90 Medical Assistants
Area Under Curve, N = 150 . . . 195
B.91 Meeting, Convention, & Event Planners
Area Under Curve, N = 125 . . . 196
B.92 Military Officer Special &
Tactical Operations Leaders, All Other
Area Under Curve, N = 130 . . . 196

B.93 Network & Computer Systems Administrators
Area Under Curve, N = 386 . . . 196
B.94 Nurse Practitioners
Area Under Curve, N = 226 . . . 196
B.95 Office Clerks, General
Area Under Curve, N = 177 . . . 196
B.96 Optometrists
Area Under Curve, N = 296 . . . 197
B.97 Payroll & Timekeeping Clerks
Area Under Curve, N = 130 . . . 197
B.98 Personal Care and Service Workers, All Other
Area Under Curve, N = 130 . . . 197
B.99 Pharmacists
Area Under Curve, N = 287 . . . 197
B.100 Physical Therapists
Area Under Curve, N = 363 . . . 197
B.101 Physician Assistants
Area Under Curve, N = 101 . . . 198
B.102 Physicians & Surgeons, All Other
Area Under Curve, N = 412 . . . 198
B.103 Police Patrol Officers
Area Under Curve, N = 106 . . . 198
B.104 Preschool Teachers, Except Special Education
Area Under Curve, N = 133 . . . 198
B.105 Probation Officers & Correctional Treatment Specialists
Area Under Curve, N = 108 . . . 198
B.106 Procurement Clerks
Area Under Curve, N = 113 . . . 199
B.107 Producers
Area Under Curve, N = 155 . . . 199
B.108 Production, Planning, & Expediting Clerks
Area Under Curve, N = 101 . . . 199
B.109 Protective Service Workers
Area Under Curve, N = 132 . . . 199
B.110 Public Relations Fundraising Managers
Area Under Curve, N = 205 . . . 199
B.111 Public Relations Specialists
Area Under Curve, N = 264 . . . 200
B.112 Purchasing Agents, Except Wholesale, Retail, & Farm Products
Area Under Curve, N = 134 . . . 200
B.113 Purchasing Managers
Area Under Curve, N = 198 . . . 200
B.114 Quality Control Systems Managers
Area Under Curve, N = 278 . . . 200
B.115 Real Estate Brokers
Area Under Curve, N = 471 . . . 200

B.116 Receptionists & Information Clerks
Area Under Curve, N = 213 201
B.117 Regulatory Affairs Managers
Area Under Curve, N = 103 201
B.118 Risk Management Specialists
Area Under Curve, N = 331 201
B.119 Roofers
Area Under Curve, N = 314 201
B.120 Secondary Teachers, Except Special & Career, Technical Education
Area Under Curve, N = 130 201
B.121 Secretaries & Administrative Assistants, Except Legal, Medical, & Executive
Area Under Curve, N = 387 202
B.122 Securities, Commodities, & Financial Services Agents
Area Under Curve, N = 183 202
B.123 Social & Human Service Assistants
Area Under Curve, N = 171 202
B.124 Social Workers, All Other
Area Under Curve, N = 493 202
B.125 Software Developers, Systems Software
Area Under Curve, N = 166 202
B.126 Software Quality Assurance Engineers & Testers
Area Under Curve, N = 166 203
B.127 Special Education Teachers, All Other
Area Under Curve, N = 115 203
B.128 Supply Chain Managers
Area Under Curve, N = 439 203
B.129 Surgeons
Area Under Curve, N = 149 203
B.130 Tax Preparers
Area Under Curve, N = 163 203
B.131 Teacher Assistants
Area Under Curve, N = 116 204
B.132 Therapists, All Other
Area Under Curve, N = 251 204
B.133 Training & Development Managers
Area Under Curve, N = 290 204
B.134 Training & Development Specialists
Area Under Curve, N = 426 204
B.135 Transportation Managers
Area Under Curve, N = 113 204
B.136 Waiters & Waitresses
Area Under Curve, N = 145 205
B.137 Web Developers
Area Under Curve, N = 192 205

B.138 Accountants & Auditors Sample
Area Under Curve, N = 981 . . . 206
B.139 Advertising Sales Agents Sample
Area Under Curve, N = 1099 . . . 206
B.140 Bookkeeping, Accounting, & Auditing Clerks Sample
Area Under Curve, N = 729 . . . 206
B.141 Chief Executives Sample
Area Under Curve, N = 2370 . . . 207
B.142 Civil Engineers Sample
Area Under Curve, N = 848 . . . 207
B.143 Community & Social Service Specialists, All Other
Area Under Curve, N = 1234 . . . 207
B.144 Computer Occupations, All Other
Area Under Curve, N = 1037 . . . 208
B.145 Computer Systems Engineers, Architects
Area Under Curve, N = 586 . . . 208
B.146 Computer User Support Specialists
Area Under Curve, N = 540 . . . 208
B.147 Construction & Related Workers, All Other
Area Under Curve, N = 2096 . . . 208
B.148 Construction Managers
Area Under Curve, N = 1015 . . . 209
B.149 Customer Service Representatives
Area Under Curve, N = 1232 . . . 209
B.150 Dental Assistants
Area Under Curve, N = 539 . . . 209
B.151 Dental Hygienists
Area Under Curve, N = 548 . . . 209
B.152 Education Training & Library Workers, All Other
Area Under Curve, N = 1333 . . . 210
B.153 Electricians
Area Under Curve, N = 572 . . . 210
B.154 Executive Secretaries & Executive Administrative Assistants
Area Under Curve, N = 958 . . . 210
B.155 Financial Analysts
Area Under Curve, N = 1317 . . . 211
B.156 Financial Specialists, All Other
Area Under Curve, N = 2072 . . . 211
B.157 General & Operations Managers
Area Under Curve, N = 2316 . . . 211
B.158 Healthcare & Support Workers, All Other
Area Under Curve, N = 2532 . . . 212
B.159 Human Resource Assistants, Except Payroll & Timekeeping
Area Under Curve, N = 782 . . . 212
B.160 Human Resources Managers
Area Under Curve, N = 1642 . . . 212
B.161 Human Resources Specialists
Area Under Curve, N = 1307 . . . 213

B.162 Information Technology Project Managers
Area Under Curve, N = 747 . . . 213
B.163 Insurance Sales Agents
Area Under Curve, N = 968 . . . 213
B.164 Lawyers
Area Under Curve, N = 1534 . . . 214
B.165 Management Analysts
Area Under Curve, N = 512 . . . 214
B.166 Managers, All Other
Area Under Curve, N = 4199 . . . 214
B.167 Market Research Analysts & Marketing Specialists
Area Under Curve, N = 1503 . . . 215
B.168 Marketing Managers
Area Under Curve, N = 1424 . . . 215
B.169 Mechanical Engineers
Area Under Curve, N = 695 . . . 215
B.170 Media & Communication Workers, All Other
Area Under Curve, N = 815 . . . 216
B.171 Office & Administrative Support Workers, All Other
Area Under Curve, N = 2559 . . . 216
B.172 Paralegals & Legal Assistants
Area Under Curve, N = 505 . . . 216
B.173 Production Workers, All Other
Area Under Curve, N = 500 . . . 217
B.174 Real Estate Sales Agents
Area Under Curve, N = 889 . . . 217
B.175 Registered Nurses
Area Under Curve, N = 1099 . . . 217
B.176 Sales Agents, Financial Services
Area Under Curve, N = 822 . . . 218
B.177 Sales & Related Workers, All Other
Area Under Curve, N = 2919 . . . 218
B.178 Sales Engineers
Area Under Curve, N = 630 . . . 218
B.179 Sales Managers
Area Under Curve, N = 1224 . . . 219
B.180 Sales Representatives, Services, All Other
Area Under Curve, N = 7749 . . . 219
B.181 Sales Representatives, Wholesale & Manufacturing, Except Technical & Science
Area Under Curve, N = 1419 . . . 219
B.182 Sales Representatives, Wholesale & Manufacturing, Technical & Science
Area Under Curve, N = 1966 . . . 220
B.183 Software Developers, Applications
Area Under Curve, N = 620 . . . 220
B.184 Transportation Workers, All Other
Area Under Curve, N = 650 . . . 220

D.1 2012 Average English US Inter-item Correlations TTI SI Style Insights . . . 227
D.2 2013 Average English US Inter-item Correlations TTI SI Style Insights . . . 228
D.3 2017 Average English US Inter-item Correlations TTI SI Style Insights . . . 229
E.1 2012 English US Corrected Item-total Correlations TTI SI Style Insights . . . 230
E.2 2013 English US Corrected Item-total Correlations TTI SI Style Insights . . . 231
E.3 2017 English US Corrected Item-total Correlations TTI SI Style Insights . . . 232
F.1 2012 Average English US α Coefficient Data TTI SI Style Insights . . . 233
F.2 2013 Average English US α Coefficient Data TTI SI Style Insights . . . 233
F.3 2017 Average English US α Coefficient Data TTI SI Style Insights . . . 233
G.1 2012 English US α-if-Item-Deleted Data Dominance,$\alpha = 0.897$ . . . 234
G.2 2012 English US α-if-Item-Deleted Data Influence,$\alpha = 0.862$ . . . 234
G.3 2012 English US α-if-Item-Deleted Data Steadiness,$\alpha = 0.864$ . . . 234
G.4 2012 English US α-if-Item-Deleted Data Compliance,$\alpha = 0.847$ . . . 235
G.5 2013 English US α-if-Item-Deleted Data Dominance,$\alpha = 0.894$ . . . 236
G.6 2013 English US α-if-Item-Deleted Data Influence,$\alpha = 0.861$ . . . 236
G.7 2013 English US α-if-Item-Deleted Data Steadiness,$\alpha = 0.861$ . . . 236
G.8 2013 English US α-if-Item-Deleted Data Compliance,$\alpha = 0.844$ . . . 237
G.9 2017 English US α-if-Item-Deleted Data Dominance,$\alpha = 0.889$ . . . 238
G.10 2017 English US α-if-Item-Deleted Data Influence,$\alpha = 0.860$ . . . 238
G.11 2017 English US α-if-Item-Deleted Data Steadiness, $\alpha = 0.854$ . . . 238
G.12 2017 English US α-if-Item-Deleted Data Compliance,$\alpha = 0.839$ . . . 239

1 Introduction

The following manual contains information on the history and development of the concept of behaviors from its beginnings to current implementation of these concepts as psychometric assessments. Included in this manual are a host of mathematical, statistical, and psychometric analyses used to establish evidence of validity and reliability of the TTI Success Insights Style Insights assessment. The introductory section contains two major parts. The first is a brief summary of some of the highlights of the company known today as TTI Success Insights. The second is a synopsis of how to navigate through this manual. A final short section summarizes how this manual is intended to be interpreted.

1.1 Executive Summary

Bill J. Bonnstetter and his son, David Bonnstetter, founded Target Training International, Ltd., in 1984 to market and sell research-based assessment solutions for the business community. With Bill's business acumen and years of experience working in the assessment industry, coupled with David's programming expertise and visionary thinking, it wasn't long before the company shifted focus from being an assessment solutions marketer to producer of the world's first personalized and computerized reports based on those assessments.

In 1986, TTI Success Insights commissioned a systematic series of independent external reviews of reliability and validity of the Styles Analysis instrument. Additional studies followed in 1992 and were published in the 1993 "The Universal Language DISC: A Reference Manual". That manual has been reprinted and updated over the years with the 18^{th} reprint occurring in 2018. There periodic updates to that original manual are now being replaced by this first edition Style Insights Technical Manual.

TTI Success Insights behavior assessment philosophical underpinnings are built on the work of such scholars as Prescott Lecky, William Marston, Walter Clarke, John Cleaver, John Geier, and many others that are detailed in this report. This pioneering evolution continues inside TTI Success Insights through continual improvement efforts in our innovative test bank review process, state of the art item and factor analysis, and cutting-edge neurological research that is exposing the actual brain activity while people take our assessments in real time.

The transition of scoring the DISC assessment from a Most/Least to a four-item forced-rank format in 2011, and a new study was commissioned that not only included U.S. based English, but also included complete reliability studies on each of the following assessment translations: Brazilian-Portuguese, Chinese-Simplified, Dutch, English-Australian, English-Canada, English-South Africa, English-UK, French, German, Hungarian, Italian, Polish, Portuguese, Russian, Spanish-Americas, Spanish-Spain, Swedish, and Turkish. A similar internal reliability review was again performed in 2017, thus maintaining this ongoing performance appraisal. Style Insights (A TTI Success Insights assessment report) is now published in more than 40 languages.

One of the primary goals at TTI Success Insights is to reveal human potential using state of the art Science of Self® assessments. TTI Success Insights is committed to continual product research and innovation, including ongoing scientific validation, new product development, and efficient product delivery enhancements.

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 family of assessments covers various tools 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 U.S. English.

Analyses of the our assessments in the many languages into which they have been adapted is reserved for a future version of this manual. There are many aspects of analyzing an assessment in a target language that present various additional challenges on top of those already posed by the baseline development of an assessment in its original language. We have attempted to provide highlights of our approach to the assessment translation and adaptation process in Appendix A following the guidance provided by the International Test Commission (ITC) in [53].

Table 1.1 presents a shortcut to the main analyses presented in this manual, signified by the blue highlighted text. The abbreviations used in the column headers for Table 1.1 are as follows. IA refers to the sections discussing item discrimination and item difficulty (item analysis or IA) which are defined in Section 4.

Table 1.1: TTI Success Insights EQ
Technical Manual Navigation

Analysis	Abbreviation	Section
Item Analysis	IA	4
Inter-item Correlation	IIC	5
Item-total Correlation	ITC	6
Internal Consistency	α, ω	7
Temporal Consistency	T/R	8
G-Theory	$E\rho^2$, Φ	9
Exploratory Factor Analysis	EFA	10
Graded Response Model	GRM	11
Thurstonian IRT	TIRT	12
Criterion Validity	N/A	13
Response Processing	N/A	14
Consequential Validity	N/A	15

IIC represents the average inter item correlation, see Section 5.1 for a detailed explanation. ITC represents the average item to total score correlation, see Section 6 for a detailed explanation. α (the Greek letter alpha) represents the α coefficient usually attributed to Cronbach, see [56]. ω (the Greek letter omega) represents McDonald's ω, see [178] for a treatment of several possible internal consistency estimates. Both concepts are presented in Section 7.

T/R refers to the section on test-retest or temporal consistency, a second part of the presentation on reliability in this manual. The third portion of reliability presented in this manual is Generalizability Theory, denoted G-Theory in Table 1.1. EFA is short for exploratory factor analysis which is covered in Section 10. GRM is the abbreviation used for graded response models, one form of item response theory particularly suited to analyzing Likert-style response data. Our treatment in Section 11 is closely aligned with the original work of Samejima in [146].

Thurtsonian IRT is a generalization of the usual item response theory modeling approach to a multi-variable item response theory model. An important byproduct of this modeling approach is that it allows for the application of the Thurstonian Law of Comparative Judgment to be applied to forced-rank data, allowing such data to be transformed into a format useable by a multi-variable item response theory model. Our approach, presented in Section 12, closely follows that of Brown, et. al., see, for example, [32, 122], among many others.

Three sections round out the remainder of the analytical portion of the manual. Section 13 presents our analysis of the relationships to other variables as defined by the American Psychological Society, see [156]. The TTI Success Insights approach to measuring response process validity evidence through gamma waves generated by electroencephalograph (EEG) analysis is presented in Section 14. This section closely follows the article [26] which presents the protocols for a similar study conducted using the TTI Success Insights EQ® assessment. Finally, Section 15 concludes the analytic sections of this manual with an outline of our approach to consequences of testing evidence of validity.

Not mentioned to this point are Sections 2 and 3. Section 2 presents an historical synopsis of the concept of behavior, along with the history and development of the TTI Success Insights Style Insights assessment. Section 3 presents our view of the current state of literature as related to the concepts of reliability and validity, as they relate to contemporary assessment development and evaluation. Finishing the manual is Section 16, which both summarizes the current body of work and puts in place the plans for future work for the next iteration of this technical manual.

A series of appendices presents information that is important to the overall discussion, yet not relevant to the current discussion. Some of this information includes historical analyses conducted over time by TTI Success Insights, see Appendices B through G. There are also appendices our approach to assessment adaptation, see Appendix A, and a presentation of our assessment development process, see Appendix H.

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

Table 1.2: TTI Success Insights Style Insights
Scale Abbreviation Table

Scale	Abbrev.
Dominance	D
Influence	I
Steadiness	S
Compliance	C

1.3 Interpretation of the Content of This Manual

As noted in the previous section, this manual is laid out based on specific assumptions and intention. The reader interested in all the details is encouraged to read the manual in its entirety. However, if the reader is interested in the final results one should bypass the main body of the manual and read Section 16 for a final view of the TTI Success Insights Style Insights assessment and the evidence of reliability and validity supporting the use of this assessment. We provide a brief synopsis of the results of the analysis presented in this manual.

The results of item analysis are presented in Section 4. These results show that the vast majority of the items fall in an acceptable range of both discrimination and difficulty based on the American Psychological Associations definition of the use of corrected item-total correlation as a proxy for item discrimination and a measure of item difficulty defined by an item endorsement approach, see the introduction to Section 4.

The results of the inter-item correlation analysis are presented in Section 5.2. Based on the analysis presented in this section, the inter-item correlations for all scales fall within the acceptable range as stated in Section 5.1. Similar analyses, presented in Section 6.2, show that the corrected item-total correlation for all items falls within acceptable ranges.

Measures of consistency are presented in Sections 7 through 9. In all cases, the scales show strong internal and temporal consistency, along with strong Generalizability and Dependability coefficients. Further, Section 7.4 presents the so-called "alpha-if-item-deleted" analysis. Generally speaking, the analysis presents solid results with regards to this measure of consistency.

Sections 10 through 12 are laying out roadmaps to the future for the TTI Success Insights continual improvement process. Some factor analytic results are included, although the item response theory results are not. The limited inclusion of results for these areas of interest is largely based on the timing of the 2020 COVID-19 pandemic. Prior to March, 2020, the TTI Success Insights research program began an extensive study of tools in our suite of assessments. Under normal circumstances, the velocity with which we are able to collect data facilitates a rather quick turnaround for research purposes.

With the onset of the pandemic and various business and community shutdowns across the U.S. and globally, our assessment collection velocity slowed to a trickle. Forced with a decision to wait out a lengthy delay in data collection or publish our current analyses, it is decided to publish our current analysis along with the roadmap to the future. This roadmap is discussed at length in Section 16.11.

The main body of this manual, presented in Sections 2 through 12, is intended to present what TTI Success Insights has done or is doing to visit the concepts of evidence of content validity and evidence of internal structure validity, the first two areas of evidence of validity noted by the American Psychological Association in [74]. The remaining three types of evidence of validity are presented in Sections 13, 14, and 15.

Section 13 presents the results to date of the measure of the Style Insights variables against external variables. The main body of external variables at our disposal are our job classification data, obtained via our demographics collection, and analyzed in this section. The interested reader is referred to both Section 13 and Section 3.3 in order to place the entire conversation into context. Section 3.3 presents an argument why the so-called validity coefficient is, at best, a weak indicator of any causal relationships that may exist between any two given variables. For further context, it is often the case that a correlation coefficient is used as a proxy for a validity coefficient.

Section 14 is a particularly interesting section as it discusses the evidence of validity based on response processing. The American Psychological Association notes the difficulty this particular area of evidence of validity presents. However, as we note both in Section 14 and in [26], the TTI Success Insights research team is working at the forefront of the combination of the neurological and psychometric, employing cutting edge techniques in an attempt to capture the response processing of the brain while responding to assessment items.

Section 15 presents an often overlooked and undervalued measure of evidence of validity. Conse-

quences of testing validity evidence is intended to show all possible consequences of the use and interpretation of assessment scores. This includes the good and the bad, with a focus on establishing a priori protocols to mitigate those possible negative consequences before they arise in practice.

As a final note, the questions of reliability and validity are not easily answered. They take a body of evidence compiled with solid information gathered over many years, in all areas of interest. In the following, we present a comprehensive look at the TTI Success Insights Style Insights assessment, while maintaining a strong focus on the future, not just the current. Our Style Insights assessment is solid and performs well in many areas. Like all assessments, continual monitoring of performance is necessary to ensure contemporary interpretations remain as valid today as they were 30 years ago when this assessment first came to life. The TTI Success Insights continual improvement process is designed to continually monitor, analyze the information obtained, and modify or replace items as necessary to maintain first class assessments.

1.4 Analyses of the TTI Success Insights Style Insights Assessment for Diagnostic Purposes

This section presents a preliminary discussion of the TTI Success Insights continual improvement process. As is noted earlier in this work, the Style Insights assessment is administered in a forced-choice format. Classical Test Theory is the most widely applied set of analytic tools for assessments, based on the ease of implementation and interpretation. These tools are applied to the forced-rank choice output earlier in this section.

It is noted, however, that Classical Test Theory tools are not directly applicable to forced-choice data. The analysis presented earlier is based on separation of the forced-choice from the ordered response string into separate scale based response data. Strictly speaking, the data, once removed from the response string, is now free from the zero determinant covariance matrix problem. This is interpreted as Classical Test Theory tools now may be applied to the data, as long as the analyst understands that the results are not completely accurate.

With this in mind, TTI Success Insights intends to implement testing strategies to gather data based on a six-category Likert-style response format in order to more completely analyze the performance of the individual items, free from the forced-choice influence of the remaining items in a given frame. The data gathering is somewhat limited, given that, in an attempt to limit test fatigue, a choice is made to gather information on no more than eight additional items appended to the end of the regular 24 frame forced-choice assessment.

It is noted that this is an imperfect approach to a challenging problem. However, at this time, TTI Success Insights is taking a diagnostic approach to analyzing our Style Insights assessment. In that spirit, we believe that incredibly valuable information is being provided by this approach. Further, this approach is planned out over the next several months, perhaps years, to be iterative and lead to a complete analysis of the Style Insights assessment.

The intent is to apply this iterative process to identify high performing items for retention, mid-range performing items for modification, and any low performing items for replacement. Once this process has identified the “best” set of 24, or possibly fewer, items for each scale, other processes may be employed to determine the “best” approach to formulating any new forced-choice frames as appropriate. These approaches may include such ideas as comparison of discrimination and difficulty parameters based on a graded response model approach, see Section 11 for a discussion, and Thurstonian item response theory models may be utilized for analysis of the multivariable item response theory models, see Section 12 for an introduction.

It is also worth noting that mathematical and statistical analysis alone cannot accomplish the end goal of this ongoing exercise. Analysis of the individual items in the face of the construct definitions and content domain must also be part of the consideration.

2 The History and Theory of DISC

The fascinating history of the DISC assessment can be broken into two epochs: Pre-Prescott Lecky and Post-Prescott Lecky. The only purpose of including the ancient history, or pre-Prescott, is to point out the over 2,500 years of effort by humans to categorize observable behavior into groupings, primarily four. The second and far more important historic era is tied to the early development of modern field of psychology and will draw heavily on the foundational thoughts of Prescott Lecky and the direct influence he had on present day DISC.

To help guide this storyline, the authors have had the rare privilege of capturing the beginnings of the theoretical underpinning in an interview with Peter Turner. Peter is the TTI Success Insights Master Distributor for the United Kingdom and was able to provide the stories that truly connect the published research to the events of the time and thus provide the underlying context. As a result, segments from that interview will be quoted directly and then built upon as necessary.

2.1 Pre-Prescot Lecky

Since time immemorial, mankind has looked inward to examine reasons and provide explanations for the world around us. We have gazed at the stars; we have focused in on the very fabrics that stitch together our universe. In these searches, we have argued and even fought to discover what it means to be human. All of us at some point have asked ourselves some form of these ultimate questions, major among them: "Who are we?" No doubt this journey began as soon as we became aware of our own existence.

The lineage of DISC can be traced to Empedocles (444 B.C.) who founded the school of medicine in Sicily. He stated that everything is made up of four "roots" or elements. These were earth, air, fire, and water. The four elements, he stated, can be combined in an infinite number of ways, just as painters can create a great many pigments with only four primary colors.

Hippocrates (400 B.C.) suggested the existence of four psychological types with his early theory of "humors". He noticed the effect of the climate and the terrain on the individual and categorized behavior and appearance for each of four climates, even suggesting which people would conquer others in battle, based on the environmental conditions in which they were raised.

Despite the fact that modern science has long departed from those four Hippocratic humors, there remains a compellingly common element regarding these Hippocratic notions. For example, our genetic material DNA is made of four nucleotides bases adenine (A), cytosine (C), guanine (G) and thymine (T). Nucleotides mix to construct genes and chromosomes with ultra-high precision in their amounts as well as sequence.

Shortly after Hippocrates' death, Aristotle took up this explorative mantle of defining our persona. He, and his student Theophrastus, developed personality sketches based on combinations of the humors using a theatrical theory. Theophrastus published 30 of these personality types in his working Characters. This particular work highlights that personalities can be differentiated and possess recognizable behaviors.

As our social and economic needs developed and grew, humans inevitably began to question many of our long-held beliefs and strived for a deeper understanding of our own existence. One of the initial among these questioners was Pythagoras of Samos. The legendary philosopher saw the world as readily explainable through logic and the application of provable thought processes. His passion for mathematics influences everything we do today. The way we process the world around us is traceable to his teachings.

This connection to psychology is not readily apparent until we discuss briefly his core philosophy. Pythagoras believed that everything within the universe could be broken down into a simple numerical form of proportions. This is key on two fronts: primarily, this is the initial basis for applying mathematics to anything. In a secondary fashion, when Pythagoras applied the question to humans themselves, he achieved a most unique answer. He determined that the human body was a separate entity from the mind (soul), essentially that thinking and acting were two parallel functions of our existence.

In a short period of time (within his own lifetime) other philosophers would take Pythagoras' theory of Dualism and study it in new and, at the time, highly exciting ways. First among these was Alcmaeon of Croton. He was known as a medicine man and considered health in perspective with this dualist philosophy. His ponderences led him to postulate that health itself consists of a state of balance between certain opposites: hot, cold, wet, dry, internal and external. Also, that not only was health due to a balance between these things, but disease itself was caused by a dominance of one of these items. From modern medicine, we see that this was as uninformed as it is ancient, however, for the time, it was an incredible breakthrough. It is interesting to note that even today the field of psychology continues to argue the existence of the mind as being separate from the brain. This dualism issue is not settled.

Moving forward, we see this theme of four structural components can also be traced to the second century Greek physician, Galen (130 A.D. - 200 A.D.), who was one of the first to describe human temperaments when he created his four-factor personality typology as sanguine, choleric, melancholic, and phlegmatic. Others classified observable human behavioral traits into far more than four groupings.

Then in 1921, Carl G. Jung (1875-1961) published Psychological Types [96]. The book was originally published in German and not until 1971 was it translated into English. This new view of typology was an attempt to reconcile the opposing theories of Sigmund Freud and Alfred Adler. He concluded that Freud's theory was based on extroverted views and Adler's on introverted judgments. Jung's book introduced the idea that each person has one of eight dominant psychological types. Jung proposed four main functions of consciousness: two were defined as perceiving or non-rational (sensation and intuition), and the other two as judging or rational functions (thinking and feeling). These functions are each modified by two overlying attitude types: extroversion and introversion. The resulting proposal suggest that the dominant function, along with the dominate attitude characterizes consciousness, while the opposite is repressed and characterizes unconscious characteristics.

Based on these definitions, the resulting types are:

1. Extraverted sensation/Introverted sensation;
2. Extraverted intuition/Introverted intuition;
3. Extraverted thinking/Introverted thinking;
4. Extraverted feeling/Introverted feeling.

While contemporary behavioral psychology finds flaws in all of these early theories, it is important to recognize these early philosophical efforts are worth revisiting in light of modern neurological understandings. The result is absolutely fascinating. For example, Jung's insight into the unconscious role of feelings and intuition is now understood as a decision-making precognitive neuro-pathway

and rational/thinking typology aligns with self-aware cognitive processing. These foundational elements continue to evolve as our understanding of neurobiology advances. More of these neurological connections to human behaviors and decision-making pathways can be explored in the following references and when appropriate, will be expanded upon in context [19], [20], [21], [22], [26], [27], [49], [48], [50], and [52]

As is true of many of these early behavioral theorists, no actual assessment accompanied their philosophic theories. For example, Jungian psychology remained for the most part remained an academic endeavor until two American women, Isabel Briggs Myers and her mother Katharine Cook Briggs, set out to find an easier way for people to employ Jung's idea that everyone has a typology. The resulting assessment is known as the Myers-Briggs Type Indicator (MBTI) and positions people in one of 16 different groupings.

It is important to note that most of the research supporting MBTI has been produced internally by the Center for Applications of Psychological Types which has raised concerns of bias and conflict of interest. Independent sources, including Robert Hogan in his 2015 book "Personality and the Fate of Organizations" called the assessment "little more than fortune cookie" [87]. Other descriptors in the literature call MBTI "Pretty much meaningless" [63], "one of the worst personality tests in existence" [3], and "the fad that won't die" [78]. Even MBTI?s webpage points out that the assessment should not be used for hiring. It is interesting to note that another popular assessment, the Big Five personality traits, shows some correlation to MBTI in four of their five scales [11].

William Moulton Marston (May 9, 1893 - May 2, 1947) can be credited with providing much of the original theoretical philosophy behind the DISC system of personality classification. Raised in Massachusetts, Marston received his education from Harvard, earning a BA in 1915, an LLB law degree in 1918, and in 1921, a PhD in psychology. Over the next ten years, Marston worked in academia, teaching at American University and Tufts University. During this time, and for years after, Marston continued his own work as a psychologist, developing his DISC Theory. Marston wrote essays in popular psychology and, in 1928, published "Emotions of Normal People" that explains the DISC Theory [120]. In this book, he describes human behaviors along two axes with actions tending to be active or passive depending on the whether the individual viewed the environment as antagonistic or favorable. By placing these axes at right angles, four quadrants are formed with each describing a behavioral style: Dominance (D), Inducement (I), Submission (S) and Compliance (C).

While the theory can be traced to Marston, it is important to notice he once again did not develop an assessment. That initial task occurred some 20 years after the publication of "Emotions of Normal People" by a former student of Marston, Walter V. Clarke. The rich story of how this occurred creates a fascinating journey involving a number of key contributors, including; Prescott Lecky, William Marston, Walter Clarke, John Cleaver, John Geier and then continues with present day tools thanks to TTI Success Insights.

Like any good story, it is best told by those most closely connected. In this case we were able to capture an interview with one of the pioneer users of the TTI Success Insights DISC tools, Peter Turner. Peter is currently the TTI Success Insights Master Distributor for the United Kingdom and, as such, was able to provide first-hand insights into the developmental path of DISC. To maintain the continuity of this interview, we will publish the conversation as it was captured and then return to key events to expand on details where necessary.

2.1.1 The History of DISC: An Interview with Peter Turner

The following interview was conducted and edited by Ron Bonnstetter. The interview took place on September 24, 2020.

> Clarke did his Master degree as a student of Marston in 1930. But he didn't appre-ciate some of the philosophical underpinnings of what Marston taught. For example, dominance is obviously a word that has sexual connotations and anyone can be influential. This takes on even more significance when you overlay the fact that Marston lived with two wives. Whilst Clarke believed that people should be able to live according to their behavioral style, he did not appreciate this particular life style. So, while Marston provides our foundation for the TTI Success Insights behavioral assessment, Clarke in particular refined and added greatly to the concept and to our understanding of what it was that we actually measure.
>
> Clarke didn't really value a lot of what Marston said. Indeed, he didn't just measure the four vectors we know he measured at least 2 others. So, Marston was seen by Clarke as very, very clever but not using the tools in ways that he thought they should be used. But the starting point for all of this, including Marston, was really tied to a guy called Prescott Lecky. I originally added him on Wikipedia. Others have taken stuff off and added other pieces about him since my original contribution, but at least he is there to be remembered. https://en.wikipedia.org/wiki/Prescott_Lecky#cite_note-4
>
> I couldn't find any reference for him, maybe, 10 years ago, perhaps when Wikipedia first came out. But now people recognize him for what he was, a truly inspirational thinker who started both Marston and Clarke by laying a philosophical foundation. He was a psychology lecturer at Columbia University from 1924 to 1934, at a time when American psychology was dominated by behaviorism. He developed the self-consistency theory as a method of psychotherapy. Self-consistency theory remains relevant to contemporary personality and clinical psychologists. Michael Stevens in 1992 captured the essence of his contributions in an article titled: *Prescott Lecky: Pioneer in consistency theory and cognitive therapy* [160]. As the story goes, Lecky asked a group of students on the first day of class, "What is the most important thing to human beings?"
>
> And majority of the students were saying things like laws, survival, food, all those kinds of things that you think of as being the basic things in life. And he said, yeah, okay. I understand what you're saying, but I've got a problem with that. So why would three-quarters of a million American servicemen volunteer to fight in the first world war? When they signed up in 1917, there were no secrets anymore. They all knew what they were signing up to do. It was over the top and many were mown down by machine gun. And then between time, it wasn't very pleasant either. And so why would they volunteer? If food, warmth, air, survival, all those things were the most important things to human being, why would anybody consider doing that? His students couldn't think of anything, so he said, I've got a theory. And my theory is, that the most important thing to human beings is not self-preservation, but preservation of the self-image.
>
> And in other words, if you believe that you should be there, if you believe that it's right, or you believe that it's a brave thing, or you believe that everything, whatever bought

your beliefs off, you will seek out the opportunity to deliver that belief. Subconsciously it's not a conscious thought, it's subconscious.

This concept actually explained a lot of things to me personally, because when Art Niemann told me this story, I finally understood, why I volunteered for impossible tasks. I remember when I was an operations guy at Halfords, the worst district in the country was under performing, it was in the Southwest and miles away from anywhere and when the job was offered, I felt my hand going up in the air, I couldn?t help it! I was drawn to be able to fix that problem. And this is what happens when people believe they are good at something. It explains why some people do things and others don?t. Now, obviously Clarke, who was sitting there in that class, thought to himself, if that's true, all I've got to do, is put some words in front of people and they will instantly be drawn to certain words. They will tick a word to say that's like them or not like them.

As a result of that experience, this became the fabric of Clarke's research that went on between 1928 to 1948. He started his own company, Activity Vector Analysis (AVA) which was part of Walter V. Clarke and Associates with the purpose of bringing his products to market. It is even more interesting to note that at the time Clarke was majoring in industrial engineering and it appears that he was in on this Lecky lecture by chance. It was also by chance he was sitting next door to Marston and these events created a calling for Clarke. But as a result of that key interaction, he changed his attention from industrial engineering to pursue a psychology master?s degree. The title of his master?s thesis was "Physical Types as Bases for Variations in Primary Emotions". So, that's kind of the basic start of all of this.

The only reason I know these stories, is because I was trained by Art Niemann and Art worked for W. Clement Stone of Combined Insurance where he was trained by Walter V. Clarke himself. The training was composed of a two-week course with Clarke, which Art said, "Changed my Life". Combined Insurance ended up purchasing Walter V. Clarke and Associates.

I remember Art telling me that it was the most powerful time of his life because all of a sudden, his life changed, as a result of understanding behavior. It made a massive impact upon him.

So that's the background of how this came about. Walter Clarke used the Marston model, but didn't like the words so he changed the words. For example, dominance first became aggressive. We know that being aggressive toward a task is okay, but also can mean being aggressive toward people, so he changed it to assertiveness. In addition to changing construct words, he created a scoring that measured both highs and lows as a continuum.

He next changed the term influence to sociability because he believed it as more descriptive. He pointed out that anybody can influence others, so it's a misnomer to think otherwise. But social people are drawn to conversations all the time, so he used the word sociability to describe what we call the "I" factor.

The "S" factor, steadiness, still creates a lot of confusion. For some it implies that a person is sloth like. If you're calm and relaxed and move about slowly, people think that's an "S" trait. Well it might be, but moving slowly has more to do with your

energy level. The "S" trait has to do with how you deal with the rate of change; the rate of change and the propensity to be calm.

Clark thought of the "S" factor as being like the waves in an ocean. When the wind blows you can be quite tormented, massive waves and people get angry. But then they calm back down like the sea that's a tranquil person's natural state. It's not tranquil all the time, but if nothing is happening, no winds are blowing, then they are very calm. They won?t go looking to make the waves and the way he described it was a bit like a swan on the surface of the water looking all calm and relaxed, but underneath their feet are paddling like mad. So, don't think that the person has nothing happening inside. Clarke says they keep it locked up. Therefore, the observable behavior of the "S" appears to be tranquil. Like the ocean without much wind.

The "C" factor was called avoidance. His first thoughts were avoiding, but again, this connotes that people avoid situations. That isn't strictly true. They avoid confrontation. So, he settled on the concept of conformance or compliance. Conscientious was also discussed, but everybody can be conscientious.

This process of finding just the right descriptive words was the result of a massive research project. Clarke was commissioned by the U.S. Air Force to interview soldiers returning from the war in 1945 and 1946. The goal was to help find jobs that would suit them. But rather than just examining their hard skills, Clarke determined and analyzed their behavioral styles. This experience provided Clarke with a deep understanding of both different styles and allowed him to refine his descriptive word choices from the tens of thousands of actual assessments plus hundreds of taped interviews and follow up conversations.

When he launched his business in 1948 he had solid information to work with and statistics to back up his ideas. I should also mention that these early assessments used a list of 81 verbs from which you checked off if they described you. Within those 81-verb list, Clarke had created not four, but five factors. So, we had the assertiveness, sociability, tranquility and conformance, but he added another one as well and that was a maturity factor.

For example, you get a salesperson who based on the first four factors, appears to be right for this job, but he says the wrong thing at the wrong time. This is an example of the maturity factor. If you had a low maturity score, then you were more likely to make these errors. A low maturity factor basically meant that you couldn't tell them anything. They were un-coachable. They used to say that the IRA terrorist were extremely mature, which at first seems to be the paradox. But what he meant was that they always felt that they were right. They knew in their mind they were absolutely right. That was the fifth factor. It might be something very similar to our (TTI Success Insights) present Emotional Quotient assessment that Clarke noticed all those years ago.

There was another factor that Clarke identified and that was honesty. He developed a mechanism by which to measure honesty, but ultimately decided it was not beneficial. This is another example of how Clarke differed from Marston. Marston believed that first and foremost, you have to tell the truth. In fact, that is the origin of the "Lasso

of Truth" in Wonder Women. It is interesting how all these things fit together. The honesty tool was never released and Art said that he never saw it.

Next, let me tell you about the AVA business. Art worked for W. Clement Stone, who founded the Combined Insurance of America (CIA). I think it is called AXON now. And I've heard lots of stories about Clement Stone, how he ran his business. I'll just give you an example of what he was like. He started the company off with, he said, $10. Well, actually the company, CIA was up for sale in the early 1920s. And he noticed it was up for sale. At the time, Clement was writing insurance policies. So, he went to the owner, I said, "I know somebody who could buy this company". And the owner said, "Oh, that's very good". Clement said, "Would you sell it, if I got somebody who could buy it?" And he said, "Oh yeah, straight away". So, Clement asked him how much he wanted for it? I don?t recall the exact amount, but it was over a million, a big sum of money even for those days. So, he said, "Right, who is the buyer?" And Clement said, "I am!" And the owner said "You can't buy it. Who is your backer?" And Clement said, "You'll back me! If you don't get any other offers, I think I might be your best opportunity."

Why? Because Clement was his best salesman. He sold lots and lots of insurance. So, on that basis, he bought the Combined Insurance of American for $10 on the basis that he would pay back the loan out of profits in the future. So that's how that whole thing was purchased.

Backing up a bit, Art had his degree in something like human resources and on the first day of work with CIA, as director of HR, he checked in with the reception desk and was told to go back to the locker room and get changed. He explained that he was the new HR director, but the receptionist insisted that his job was to take the mail around to offices. So, he delivered all the mail to people for about three or four weeks. One day he bumped into Mr. Stone and asked "When am I going to start working at my director of HR job?" Mr. Stone says, "You are already! You've met everybody, you've talked with everybody when you take them coffee. I want you to continue to do that." He says, "And this is your job description." He turned piece of paper over on his desk and said, "Make my people happy". That was his job description, "make my people happy". So, he continued to go around fixing the lights interacting with people.

What he learned from all that was that it isn't the big things that upset people in work. It's the little things. And by understanding how people behaved and what their expectations were, he was able to make people happy. All this tied back to Clement Stone. As a result of his background with AVA, he actually commissioned them to do some research for him later on. And I found a paper today, which shows that Clarke was doing this kind of analysis for insurance sales people.

There is a white paper written on that somewhere, which you can find. And so that's the result of that. He got Art to go and be trained by AVA. So that's how all that kind of ties in with Clement Stone. Now, Clement Stone in his books, wrote in there saying that the only way that his business grew from being a $10 company to a multi multimillion-dollar company was by the use of behavior profiles; making people happy by them doing the job which suits their natural behavior. So, he credited the whole of his growth to being able to use these tools to first of all, identify what the job was and

then match behavioral styles to that job. They used a tool called JAR (Job Activity Report) to define the job, which is the same as our work environment assessment.

It is important to note that the overall ethics of the AVA company, the Walter V. Clarke company, was that they were there to do good. That was the philosophy they built the company around. This was a very, very powerful vision. You don't pigeonhole people, you don't have highs and lows. You don't have good and bad styles. You just create this continuum of styles that allows you to understand somebody's preferred style of behavior. Each of these shapes had a name. So, you didn?t really pigeonhole people but you did place them within this continuum so you could understand where somebody was and use this to match them to appropriate jobs.

J. P. Cleaver, "Clipper Cleaver" as he was called, joined Clarke in 1951 and left in 1956 to start his own company. Now according to Art Neiman, Cleaver took all of Clarke's research, when he left. He took a cabinet of papers with all his research. You would think that that would cause major upset, however, within two-or-three years Cleaver came back to Clarke and asked him to help him to devise a new instrument. So, Clarke worked with Cleaver and came up with a closed questionnaire instead of the list of descriptive words Clarke had created. These are the kinds of questions that we see now and in the format that we use.

The result of this collaboration was a second alternative approach. In August of 1963, Walter V. Clarke published a paper titled: *Forced choice versus free response in personality assessments*. That's quite an interesting paper, especially when you consider that Clarke continued to use the open question format. And I think they do to this day. J.P. Cleaver, "Clipper Cleaver", went on and used the closed response instrument. So, Clarke was involved in both assessments. Cleaver also turned the graph so they were vertical and removed the maturity factor, and went back to using the Marston's DISC nomenclature that we see today.

So, Clarke had his hand in these new assessments, but he didn't agree with the use of highs and lows and, therefore, stayed with his original open-question approach. I should add that there never were any patents filed on anything or copyrights because it basically belonged to Clarke and he didn't object to their use. Clarke actually helped JP Cleaver grow his business because it was for the good of humanity, as he saw it. So, a very strong ethics drove this early work.

Going on from that, Art Neiman started his own company in 1980. And about 12 months after that Mr. Stone's Combined Insurance of America bought the AVA company. AVA subsequently has been sold again to somebody else. I spoke to the new owner once, but can't remember who they are right now.

Going back to Clarke's open question assessment, the number of verbs that you ticked actually created what he called an energy line. So, in other words, if you ticked a lot of verbs, you were said to have bags of energy. If you only ticked a few, you didn?t have much energy. As far as I know, there is absolutely no research to support this conclusion. I've never seen anything to say that that is a valid way of measuring if someone has any energy or not. Whatever energy is any way. I don't know, it might be more accurate to measure if you had a pint of beer or two the night before.

> In preparation for this interview, I found more information, white papers and other publications. One in particular is a good description of early personality measurements. This 1965 paper was published in the Journal of Clinical Psychology by Clarke, and a guy called Peter Merenda [127]. Now I know about Peter Merenda because Art told me about him. He worked with Clarke for a while and earlier to today I found out he only died about 12 months ago. He would have been able to tell us lots of things about all this. I did learn that in 1957 he got his Ph.D. from the University of Wisconsin in Madison and he was assigned to the Naval Examining Board.
>
> He partnered with Walter Clarke and was a professor of psychology at the University of Rhode Island where he co-founded the Department of Psychology and Computer Science during his tenure in academia. He was one of the original members of the American Psychological Association and was granted the prestigious Samuel J. Massey award. Over the years, Dr. Merenda developed close relationships with scholars all across the globe. He and Clarke wrote loads of papers with him. For example, another one I've got here is about the construction of an industrial selection personality test. That's riveting reading by the way. That was published with Walter Clarke in 1956 [40]. This connection with academia adds a lot of credibility to Clarke's work.

2.1.2 Expanding on Peter's Interview and Highlighting the Role of Prescott Lecky

Until Peter Turner mentioned a significant contribution of Mr. Lecky to the history of DISC we, for the most part, had never mentioned him. Upon further investigation, an exciting set of foundational components were discovered that in many ways lay a much more solid theoretical foundation of DISC than Marston provides.

We learned that Lecky pioneered the concept of self-help as a method of psychotherapy. This new theory was built on the belief that the most important motivator for humans was the preservation of the self-image. He understood that one's self concept dictated their resulting behavior. His transforming concepts directly influenced Maxwell Maltz's classic self-help book, "Psycho-Cybernetics" as well as George Kelly's book, "The Psychology of Personal Constructs". In addition, Lecky inspired several notable students, including Hans Ansbacher, Walter Clarke, Gardner Murphy, Frederick Thorne, and Carl Rogers [4]. During the winter of 1927-1928, Lecky studied with Adler in Vienna, visited Jung, and, possibly, met Freud during that same trip [167].

In the 1990s, Michael J. Stevens traced the foundational elements of numerous cognitive-phenomenological theories back to Lecky's by preforming a massive search of the literature as well as numerous interviews. His findings were first presented at meetings of the American Psychological Association in 1990 and again in 1991. These findings were then published in the Journal of Clinical Psychology in 1992 [160]. A quote from that paper starts to set the stage for understanding just how influential Prescott Lecky has been and how his work continues to drive our theoretical understandings.

> Self-consistency theory represents a psychosocial extension of the gestalt motive for physical order and constancy (Swann, 1990). As such, Lecky anticipated several phenomenological and consistency theories, including Combs and Snygg's self-theory, Rogers' person-centered theory, Kelly's personal constructs theory, Secord and Backman's interpersonal congruence theory, and, more recently, Epstein's cognitive-experiential self-theory, Swann's self-verification theory, Higgins' self-discrepancy theory, and Andrews' self-confirmation model. Historical and situational elements of Lecky's theory, particularly striving for unity, are less well-represented in Festinger's theory of cognitive

> dissonance, Heider's balance model, Murphy's perceptual learning process, and Osgood and Tannenbaum's congruity principle. (pg. 808)

So why are we not more aware of this man's turning point influences? The reasons may lie at the heart of his own DISC profile. Imagine a man who has an off scale theoretical motivator and a major "D-C" (potential me-me conflict). Now overlay this profile with the fact that his position as psychology lecturer at Columbia University was contingent upon the completion of his dissertation and as such he was given a ten-year window to complete this task.

Lecky's former student, John Francis Adams Tayler sums up such a scenario when he states in a biography about his mentor:

> Although he had accumulated sufficient data for an acceptable thesis at several points in his career, he could never bring himself to completing the final details necessary for the accomplishment. Part of this inefficiency may be attributed to perfectionistic tendencies which made him unwilling to publish anything which did not reflect his latest views. He revised his dissertation again and again, but was never completely satisfied with the results.

At the time of his death, Lecky had only produced two professional paper presentations [107] and [108], and a book review [106]. However, he left a collection of incomplete manuscripts and a handful of former students who were determined to eventually make known his name and his Self-Consistency theory of personality work. As a result of this group determination, in 1945 a 154-page book was posthumously published under his name in compliance with government regulations on conservation of essential materials [109]. In 1951 the original manuscript was expanded to 275 pages with interpretations added by Frederick C. Thorne [110]. Thorne was Lecky's graduate assistant from 1938-1939 and a critical admirer of Lecky as a scholar, teacher, and clinician. This 2nd edition was reprinted again in 1961 [111].

2.2 Walter Vernon Clarke

Armed with the combined influence of Marston's Emotions of Normal People and Lecky's theory of self-consistency, Clarke was the first to build a test for personnel selection called Activity Vector Analysis in 1942. Clarke identified a list of descriptive adjectives and using a check list format, he asked participants to check the specific words that described them. The entire process that followed to create this new style assessment is extremely well documented in his 1956 publication titled: *The construction of an industrial selection personality test* [40].

Using state of the art psychological tools of the time, Clarke followed the Lexical approach, which is based on the assumption that the most important personality traits are encoded as words in natural languages and that the analysis of the structure of those words may lead to a scientifically acceptable personality model. It is important to expose the rigor employed in this early assessment development for the reader to understand not only the laborious process followed but the necessity of rigor in assessment development and continual improvement processes. With that goal in mind, let us outline some of the steps taken by Clarke to ensure credibility and to establish transparency. This information is being provided to not only highlight the rich historical development of behavior assessments, but to expose how the remainder of this technical manual takes these initial developmental processes to the next level.

2.3 Understanding the Assessment Creation Process

2.3.1 Development through Pilots

Clarke in his 1956 paper details each step in his assessment developmental process, starting with how he identified a list of 183 adjectives that were commonly used in describing others. He collected information on the adjectives using a checklist format and asked applicants for jobs at a large machine manufacturing company to check the specific words that described them. After tabulating the number of persons checking each word, words were eliminated if less than 5% responded or more that 95% of the participants chose that adjective. After several refinement pilots, a final list of 81 words met all of the standards applied. After collecting and analyzing the data on this instrument, he discovered the four factors produced from the data (aggressive, sociable, stable, and avoidant).

2.3.2 Weighted Scoring Explored

Weighted scores were correlated with unweighted scores with $r = 0.93 \pm 0.01$, therefore indicating little added accuracy of interpretation was gained by weighting, so the process was not employed further.

2.3.3 Initial Norms

Preliminary norms were established from the first 100 cases in 1945 and later 500 cases were randomly selected from 1067 responders.

2.3.4 Item Re-analysis

In 1947 each word choice was verified to correlate with each of the four vectors. Three words were eliminated from the list because of correlating less than 0.30 with the vector score. An additional 75 words were piloted and three substitute words were selected as replacements.

2.3.5 Standardization Step

When over 15,000 tests had been accumulated, norms were reestablished.

2.3.6 Split-half Reliability

Coefficients of internal consistency were found for each of the four constructs to ranging from 0.92 to 0.97 ± 0.02.

2.3.7 Test-retest Reliability

Test-retest reliability was determined for 323 varied occupational cases with a mean retest interval of one year (minimum one month) contributed by a large number of business and industrial companies located in various geographical areas of the United States. Coefficient of stability for each construct ranged from 0.62 to 0.75 ± 0.03.

2.3.8 Validation (Similarity of Pattern)

Clarke states, "Validity of the analysis would be indicated if it could be shown that significant differences exist between the personality patterns of persons successful on one kind of work, as compared to persons successful on work of a different nature." In this initial paper, Clarke compares

persons classified as President, Executive, Salesman, Accountant, Engineer, Teacher, and Machine Operator. Clarke continues his rationale for comparing these job groupings, by stating, "If this instrument is a useful tool for selecting salesmen, executives, accountants, or the like, there must be some significant differences among the patterns obtained from these people," pp. 384. The study was able to show that "fairly significant differences" were found between most of the jobs and where differences were not found, similarities between the requirements of the jobs appeared to justify the lack of differentiation.

It is interesting to note that from this initial reporting of the assessment construction, Clarke generated a series of articles that each described the unique qualities identified by his assessment tool. These follow up articles expand on how these tools can profile: life insurance agents [41], loan office managers [42], self-made company presidents [43], prediction of occupational hierarchy [44], AVA as a selector of life insurance salesmen [171]. Upon release of this documentation and protocols, a host of other predictive publications followed by other authors [88] and [98].

2.4 John Cleaver: From AVA to DISC

As stated in the Peter Turner interview, John P. Cleaver initially worked for Walter Clarke and Associates and in 1956 left and established J.P. Cleaver Company. This first assessment tool was called Self-Description and instead of focusing on one's relationship to a job, this tool focused on one's behavioral characteristics in general. The initial tool was similar to Activity Vector Analysis as an adjective checklist, but after turning once again to Walter Clarke for advice, it evolved into 24 sets of 4 words (96 adjectives) and required that the participant select the adjective most like and the one least descriptive in this forced-choice instrument. Scoring was accomplished using a layover template and then visualized using three graphs. Factor analyses of the Self-Description produced two factors that closely approximated the underlying axes of Marston's model, lending considerable empirical support not only to the structure of the model he proposed, but to Clarke's earlier claim that an adjective-based instrument could be created. Thus in 1965 we find the first reference to the DISC assessment tool. In addition to this early DISC assessment, Cleaver also introduced the "Human" Job Factors survey of 24 statements. Participants responded on a 5-point scale to indicate the importance of each in the successful performance of a job.

2.5 John G. Geier's Contribution

John G. Geier, Ph.D., Dorothy Donney, Sam Gardiner and Mike Johnson, Ph.D. were all involved in the early business development of Performax Systems International, Inc. Bob Picha and Betty Bowman were involved in early training and marketing. However, Geier (January 24, 1934 - September 26, 2009) was the leading force behind the business. The publication of "Energetics of Personality: Defining a Self" and "Personality Analysis", both in 1989, solidified his contribution to behavioral assessments [71] and [72]. The company implemented a multi-level marketing approach in the late 1970s and Performax Systems International, Inc. was sold to Carlson Marketing Group in 1981. Carlson was later sold to Groupe Aeroplan in 2009.

Many are aware of Carlson's DISC products; however, not very many people are aware of the historical development of these DISC products. John P. Geier developed the Personal Profile System™ in 1963. The first instrument was only 4 pages long and included the responses, score and graphs. All feedback came from a person trained to interpret the graphs. The instrument used carbon to transfer the responses to the scoring page. The original instrument used 96 adjectives (24 boxes) and had three graphs. The adjectives and graphs were very similar to J.P. Cleaver's

DISC instrument developed in the 1950s.

Later, Geier added 15 classical graphic layouts to the instrument. Users had to eyeball their graphs and look at the classical for similar graphs to find the correct feedback. Later, segment numbers were added to the graphs to provide a system to look up the correct classical layout for feedback. The Personal Profile System™ places an emphasis on Graph III (the average graph).

Over the years their opinions have varied in regard to which of the three graphs is the most important. Early on they had used and trained on how to use all three graphs. Then they appeared to de-emphasize Graph III and placed this emphasis on Graph I and II. More recently they are advising users that Graph III is the most valid.

By comparing their early works (Cleaver with John Geier's Personal Profile System), one can see many similarities in both the 96 adjectives and the 24 Job Factor Statements. Geier's profiles appeared in the 1970s. At this point it should be noted that many versions of behavioral assessments based on the four-factor model were developed from the 1960s up to 2000 and that number continue to expand.

2.6 The Creation and Expansion of Target Training International, Inc.

Bill J. Bonnstetter (January 22, 1938 - June 2, 2016) also is considered one of the pioneers in the assessment industry. Bill attended the Iowa State Teachers' College, where he earned his bachelor's degree in business with an emphasis in marketing in 1964, then immediately went on to earn a master's degree in business education, in 1969, from the University of Northern Iowa. Driven by a passion to help others better understand self, Bill connected with Carlson Marketing Group in the late 1970s.

He immediately started to collect his own research with an understanding that doing one's own research allows for insights into an assessment tool that can be gained in no other way. Because of his background and ongoing interest in sales, Bill conducted a farmer buying style study. This involved having midwestern farmers take the DISC assessment and then systematically capturing photographic evidence of each farmstead. Images were taken of their farm building, major equipment, the lane leading to their property and any other visible component that might relate to the farmers behavioral styles. From this research, Bill was able to match visual farmstead cues to primary DISC behavioral styles [16].

In 1984 Bill's son, David Bonnstetter, added his computer skills to the business that resulted in the first computer delivered DISC behavioral assessment and a personalized report. That same year Bill and David co-founded TTI Success Insights.

Early on, Bill understood the value of positioning one's self in the market place by filing for patents on inventions that had the potential of revolutionizing the industry. His first patent was filed with colleague Jon Hall in April, 1995, under the title, "Employee success prediction system" and was granted in September, 1996, with the US Patent No. 5,551,880 [163]. In January, 2001, TTI Success Insights filed a patent application for the web based "Network based document distribution method" was issued US Patent, No. 7,249,372 [164]. These two patents were quickly followed by a third in 2007 that documented a method for first deriving key characteristics for superior performance in a job by identifying a set of behavioral-related competencies relevant to specific jobs and then surveying applicants to determine the best candidates [165].

Even before these cornerstone patents, TTI Success Insights established and ongoing research agenda that has grown and matured over the years. Starting in 1986, TTI Success Insights com-

missioned a systematic series of independent external reviews of reliability and validity on the Style Insights instrument. The first report was performed by Professor Russell Watson from Wheaton College in 1986 and repeated again in 1989. Additional studies followed in 1992 and were published in "The Universal Language DISC: A Reference Manual" in 1993 [173]. That manual was reprinted and updated again by Warburton and Suiter in 1999. This revised edition included an updated study by Associate Professor James R. Hall, Chair of the Department of Psychiatry & Human Behavior at the University of North Texas, Denton, Texas [174].

With the transition of scoring the DISC assessment from Most/Least to using all four item rankings in 2011, a new study was commissioned. Once a large enough sample size was collected using the new scoring protocols, the next external review was expanded in 2015 when Professor Delwyn Harnish, Director of Assessment at the University of Nebraska-Lincoln, ran reliability studies on not just US English, but preformed complete reliability studies on each of the following assessment translations including: Brazilian-Portuguese, Chinese-Simplified, Dutch, English-Australian, English-Canada, English-South Africa, English-UK, French, German, Hungarian, Italian, Polish, Portuguese, Russian, Spanish-Americas, Spanish-Spain, Swedish, and Turkish [85]. A similar internal reliability review was once again performed in 2017, [166] thus maintaining this ongoing performance appraisal. With Style Insights (a TTI Success Insights DISC assessment report) now published in over 40 languages, TTI Success Insights will continue to analyze each of these translations, once a large enough sample size can be collected.

2.7 Returning to the Issue of Average Graph Values

In 1984, TTI Success Insights identified Graph III (average DISC scores) as not providing valid information, especially when there is a significant disparity between Graph I (adapted) and Graph II (natural). Figure 2.1 provides an example using data from one of the authors. At the time of this report, this individual was unhappy with his university position and was considering a professional move.

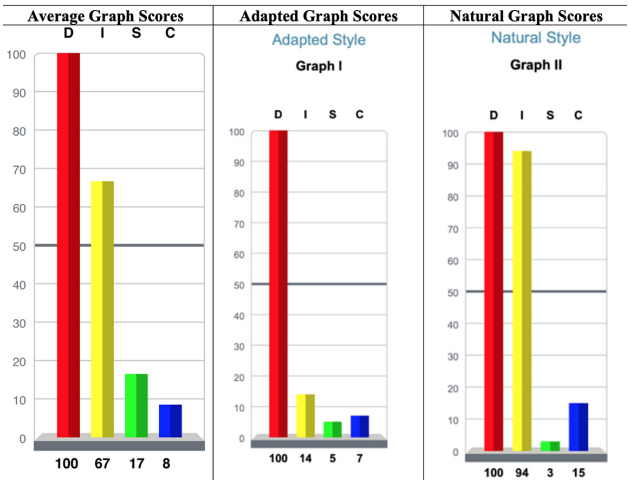

Figure 2.1: Average, Adapted, & Natural Graph Comparison

Take a moment and think about debriefing the information provided from the far-left graph, without either of the other graphs. No doubt the debrief would focus on the dominant nature of the "D" factor, the rather low "C" and the "I" score that is well above the midpoint. Now add to the story by including information from the middle (adapted) graph. One immediately sees that this person

is adapting their "I" factor down rather dramatically. The final piece of the puzzle only emerges when the natural style is exposed in the far-right column. The natural scores are generated by using what a person says they ?are not?. The significance of this negation will be expanded upon in a moment.

While the original research references to support this decision has been lost with the passing of Bill Bonnstetter, anyone who has debriefed numerous reports can recall scenarios when critical issues only emerge when comparing the natural and adapted, such as those in Figure 2.1.

One of the clearest historical references in support of this dual world view is found in an article by Dr. David Warburton, [172] titled: *The importance of finding the real person at work*. Please note that he does not say the importance of finding the "right" person, but the importance of finding the "real" person. Warburton's paper lays the groundwork for these natural and adapted behavioral difference when he states:

> All of us have developed behavioral styles, which are represented as distinct cognitions, emotions and actions. The central core of our behavioral style tends to remain stable because it reflects our individual identities. However, the demands of the work setting often require different responses that evolve into a work behavioral style. In order to help people understand themselves, we need to evaluate this disparity. (pp. 2)

Warburton goes on to explain how this separation of perspectives requires an understanding of how people think about themselves. He grounds his comments in the work of Lamiell [102], [103], and [95], who points out that human judgements are not framed by contrasting themselves with others, but in fact frame their self-identity by negation [37]. This literature states that our identity emerges through a process of articulating what we are not, not what we are. To make this decision-making process even more fascinating, is that much of this mental processing draws upon implicit mental processes for which we are not even cognitively aware of at the time. As a result, self-assessment ratings that expose what a person is not, provides a more direct path to the person's true view of him/herself [172].

Warburton and Suiter [174], further exposed the value of this distinction between Graph II (Core Behavioral Style) and Graph I (Responses to Environment) in a study of 150 middle managers. At this point, the TTI Success Insights DISC assessment required the participants to simply select the one item that was most like them and the one item that was the least like them, with the least being assigned to their natural or core style and the most representing their adapted or response to environment. (Notice how once again what one is NOT becomes the natural or core style.) The study found that differences in the two DISC graphs were a predictor of problems at work and at home. Furthermore, the distinction between acceptance and rejection was found to relate to the amount of alcohol use, physical health, mental health, job satisfaction and absenteeism. The paper goes on to point out the immense physical and financial costs of job dissatisfaction that can be exposed when examining these discrepancies.

While these findings speak volumes for the need for two different graphs, research always seeks verification from different approaches to the same issue. The verification came about as a result of a unique assessment improvement protocol that TTI Success Insights was the first to employ. In 2011, Ron Bonnstetter and Dustin Hebets began collecting real time brain imaging data while participants took the TTI Success Insights DISC assessment. The basic process in depicted in Figure 2.2. The key to understanding this line of research that tied to the fact that electroencephalogram (EEG) assessment data showed totally different brain activity between accepting a descriptive word compared to rejection of descriptive word or phrases.

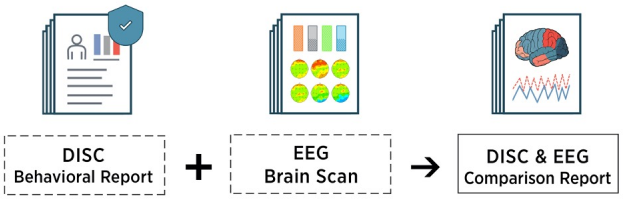

Figure 2.2: Neurological Support for Two Graphs

Figure 2.3 illustrates three basic different participant responses to a DISC forced-choice assessment cell. Notice how the intense red gamma activity is dominate on the left side when responding to a positive reaction and to the right side when rejecting a concept. (Point of clarification, these images are showing the brain as it would be looking at you, so the left brain is on the right side.)

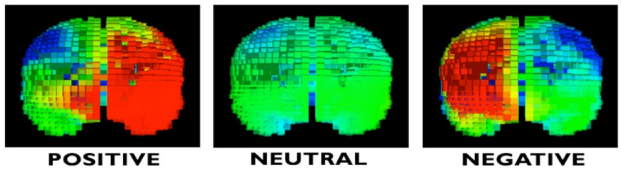

Figure 2.3: S-Loreta Image of Gamma Activity in the Frontal Lobes

The images used in Figure 2.3 showing acceptance and rejection sample images were included in our approved United States patent application. In other words, the patent office has affirmed our ability to differentiate these two ends of the choice spectrum by issuing, "Validation Process for Ipsative Assessments", US Patent No. 9,060,702 [17]. This original patent was reaffirmed when a similar patent was file and approved by Canada [18].

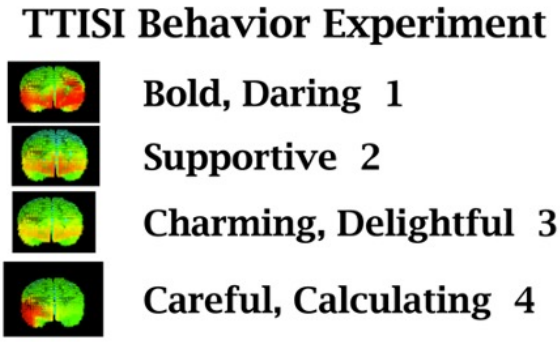

Figure 2.4: TTI Success Insights Experiment

Figure 2.4 shows how this process has been used to both validate assessment items and to gain additional knowledge regarding participant decision-making. This individual selected ?Bold, Daring? as the most descriptive word or phrase and "Careful, Calculating" as their least descriptive terms. In other words, the person "accepted" item 1 and "rejected" item 4. Notice how the red area of the frontal lobes moves from a dominate left side of the brain for their 1^{st} choice and clearly to the right side for their 4^{th} choice. Clearly there is a difference in how the brain processes this data and TTI Success Insights has collected an abundance of data to support this difference.

In summary, it is important to note that if responses to the acceptance rankings, (1 and 2) are

very similar to the DISC scores derived from the rejection rankings, (3 and 4), then the two graphs would also be very similar. A case can be made that having two graphs that are similar is a strong indicator that the persons core behavior and their adapted behavioral styles are in harmony. Those DISC providers who only offer one averaged graph would find affirmation in this scenario. However, when the two graphs are different, an explanation needs to be explored and both the historical findings and the neurological data clearly make a case for this additional detailing.

Simply put, DISC behaviors cannot accurately be described using one graph when in fact each of us make decisions that are based on an initial, precognitive, neurological asymmetry that draws upon totally different brain processing of acceptance and rejection. Forced-choice data in a single graph is analogous to comparing apples and oranges as one thing.

The human brain starts a decision by grouping input along a continuum that is documentable in the frontal lobe asymmetry. In addition, and this is a crucial statement, humans process rejection input faster and with greater intensity than acceptance! This trait has been instrumental in our survival as a species. It also means that what we say we are NOT is a stronger reaction and a clearer indicator than what we say we ARE. Therefore, scoring a person?s 3^{rd} and 4^{th} choices (least descriptive) actually provided an inverse indicator of who we really are, thus our natural style comes from items 3 and 4. Our adapted style then is secondarily processed and provided by scoring items 1 and 2 as an indication of our adapted behavior.

It is always rewarding when observed findings by early researchers, such as Warburton, Suiter, and Bill Bonnstetter, are later explained by more recent neurophenomenological findings. I could provide much more about the role of the emotional load of the item and other details, but the references below can provide those exciting details that I am sure you will want to embrace.

Since the filing and approval of these original patents, TTI Success Insights has successfully published over a dozen peer reviewed papers that use this neurophenomenological gamma asymmetry to document acceptance versus rejection [17], [18], [19], [20], [21], [22], [26], [27], [48], [49], [50], and [52].

2.8 History of DISC Concluding Comments

The TTI Success Insights behavior assessment has come a long way from the initial Marston theoretical influence. As you can see, many others made major contributions along the way and created this ongoing evolution of behavioral understanding. This evolution continues inside TTI Success Insights through continuous improvement efforts that include our innovative test bank, state of the art item and factor analysis and cutting-edge neurological research that is exposing the actual brain activity while people take the assessments in real time.

In the words of Peter Turner:

> There's no question about it. I've never once in all the thousands of reports I have personally fed back to people, never once have I had reason to doubt what was reported. I suppose that's a kind of self-reporting face validity. And I have only had two people tell me the report was wrong. And one of those was a person who said, it's wrong that you should follow me around. And to find out how I live my life. He really thought that I had written the report myself by following him around.

The history of DISC is rich and the future is even more exciting as we explore new ways to improve on these foundational efforts through continued research and development.

3 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 [74], [75], and [76]. 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 [5]. 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 [156], 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.

3.1 Thoughts on Evidence of Reliability

We go more in-depth into the history of reliability coefficients in the section on internal consistency in [25]. 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 [101]. 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 [38] and [86]. [38] is a bit more useful in that it discusses several different approaches and offers some suggestions as to directions one could take. [86] is a nice exposition on an empirical analysis of the use, or misuse and misunderstanding, of the α coefficient specifically.

[38] 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, [86] 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 [38] and [86], 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 [38] and [86] present slightly differing views, they generally agree that [101] first proposed a "reliability formula (called KR-20) that can be used for data on dichotomously scored items...", see [38]. [86] states that "Alpha was introduced by Kuder and

Richardson (1937) for dichotomously scored items". We point out this difference given that [38] does not refer the result of [101] as "alpha" while [86] 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.

[38] traces the history of reliability coefficients to Spearman and Brown, both independently publishing in the same issue of the *British Journal of Psychology*, see [158] and [34]. The formula published has since become known as the Spearman-Brown formula. [86] begins their discussion with [101]. 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 [38] is introducing.

Table 3.1: 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 [38] was published in 2015 and [86] 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 3.1, above, is presented in [59] and reproduced here for purposes of the current discussion. Additionally we present the following quote from [10]:

> 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 [131] stating that 0.70 is acceptable during development, 0.80 for "basic research scales", and 0.90 for clinical settings are appropriate minimums. [142] recommends a minimum of 0.80 without the same restrictions as [131]. As a final example, [97] suggests values of 0.95 for clinical settings. [10] 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?

3.2 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 approaches 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 intended to really 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 [74] 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 [156] 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.

[5] 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 [5] 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 [5], 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 [156].

> 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 [26].

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 [156] 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 [156] 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, [25]. 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 [25]. 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 [156]. Also discussed are IRT models, multidimensional scaling, and evaluating the invariance of the assessment structure through the use of differential item functioning. This last area is an important one in the US for legal reasons. Differential item functioning 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 Thurstonian 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 3.5 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 [5], 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 3.3, 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 [156] 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 [156], 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?

3.3 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 3.1. 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.

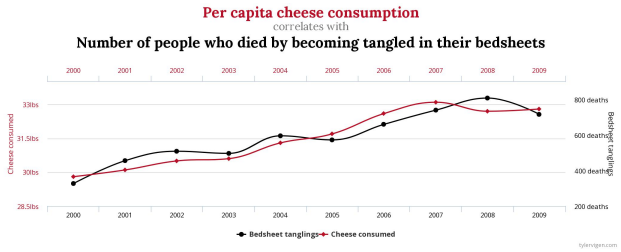

Figure 3.1: Bad Correlation v Causation Example

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.

3.4 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 [153]:

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. [12], [118], and [135] 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 [12], 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 [6], [9], [65], and [176] 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.

3.5 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 3.2 and presented in more detail in section on Thurstonian IRT models in [25].

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 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 [94], 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 [94] 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 [94] 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 [94]. 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 [94]. The work presented in [30, 31, 32, 122, 123], 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.

3.6 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 [93]. 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.

4 Item Difficulty and Discrimination

Item analysis is a widely encompassing and quite simple to implement analytic tool that provides a great deal of information about both reliability and validity of psychometric assessments. This tool provides powerful information at all stages of development and use of assessments. In particular, item analytic approaches may be used to identify both strong and suspect items during assessment development. Additionally, periodic review of item analyses provides invaluable information about the performance of the assessment and is a critical part of any continual improvement process.

This section is broken down as follows. First, a generic description of item analyses is given following that provided in [136]. The authors in [136] approach item analysis from the standpoint of Item Response Theory (IRT). As such, [136] presents an IRT and corrected item-total correlation based approach to item analysis. This does not pose any problems for the current analysis as the methodology employed in [136] is merely a generalization of certain ideas presented in this manual. Further, Sections 11 presents an IRT approach based on the concepts behind the Graded Response Model and Section 12 presents a view of the Thurstonian Law of Comparative Judgment as presented by Brown, et. al., see any of [30], [31], [32], [122], and [123], among many other possibilities.

We introduce the concept of item difficulty from the standpoint of respondent endorsement. [136] presents item discrimination as being approximated by corrected item-total correlation based on a given population. TTI Success Insights has adopted the American Psychological Association's (APA) approach from [136] for item discrimination and item difficulty in this section with generalizations presented in Sections 11 and 12. In other words, our item analysis uses corrected item-to-total correlation as an approximation for item discrimination and one minus the percentage of positive endorsement as an approximation for item difficulty.

4.1 A Short Review of the American Psychological Association's View of Item Analysis

As alluded to in the previous section, this manual routinely utilizes the information found in [74]. In particular, this section relies on the results of Chapter 7 of that work, Item Analysis, see [136]. When considering item analysis it is quite helpful to keep the following lengthy, but highly important, quote in mind. The bold face text is added by the authors of this manual for emphasis.

> The properties of item discrimination and difficulty play **fundamental roles** in determining the **reliability** and **validity** of the final scores generated by the assessment. Without adequate discrimination across the items of the assessment, the assessment cannot generate scores that are valid and reliable, regardless of the target trait levels of the individuals in the population. Without an appropriate level of difficulty across the items of the assessment, the assessment cannot generate scores that are valid and reliable at particular points of the target trait continuum.

If one defines the psychological trait of interest for a given psychometric assessment to be the target trait, item discrimination is a measure of how well the response categories of an assessment differentiate between individuals at different points along the target trait continuum. In contrast, item difficulty measures the target trait level at which the information provided by the item is at its maximum. This concept may be better explained through the use of an example.

Consider the information presented in Table 4.1. This table contains a fictitious example comparing the level of information provided by a set of items on an assessment and their difficulty, or ability to

measure the target trait. Figure 4.1 presents the data from Table 4.1 as a scatter plot of Information v. Target Trait. The interpretation is as follows. Point A shows that item A provides a moderate level of information at a low target trait level or low level of difficulty. Contrast this

Table 4.1: TTI Success Insights
Information v. Trait Example

Item	Target Trait	Information
A	0.15	0.40
B	0.50	0.80
C	0.65	0.05
D	0.90	0.40

with point D which shows a similar level of information at a high target trait level. Points B and C show high and low levels of information at moderate levels of the target trait, respectively. Level of target trait indicates the position along the target trait continuum at which information about an individual is provided. The height of the circle indicates the amount of information about the individual that is provided at a given level of target trait.

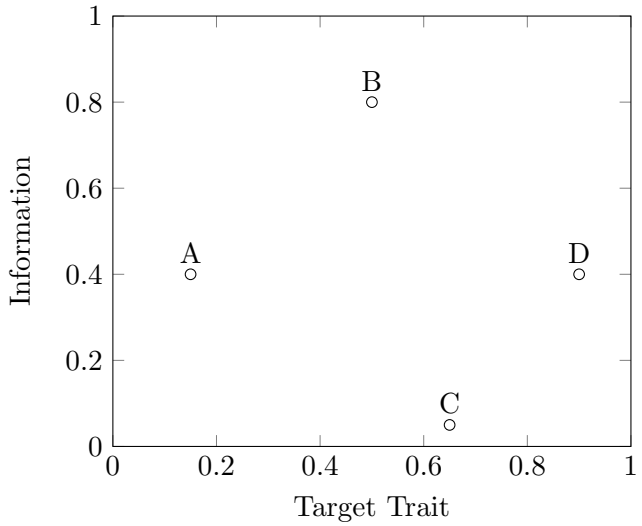

Figure 4.1: Information v. Target Trait

4.2 TTI Success Insights Style Insights Forced-rank Item Analysis Results

Based on the previous discussion, we desire scales that cover a broad range of the trait measured by the scale with higher discrimination values desired over lower ones. In other words, the more discriminating an item is at a given difficulty level, the better that item is at discriminating individual respondents at that given level of difficulty.

With the previous discussion in mind, all four scales appear to cover broad ranges of item difficulty at lower to moderate ranges of discrimination. In an ideal world, one may desire higher levels of discrimination generally speaking. Given that the current analyses is completely diagnostic

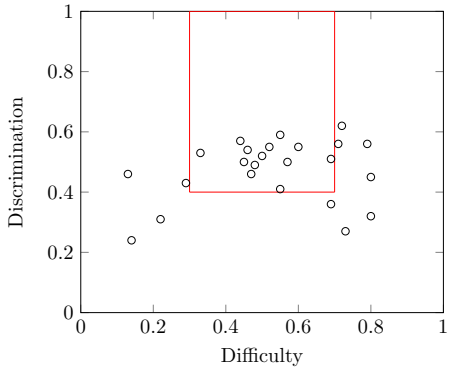

Figure 4.2: Forced-rank Dominance

in nature, a major focus of the future development of the TTI Success Insights Style Insights assessment is on increasing the levels of discrimination while maintaining broad coverage along the trait continuum.

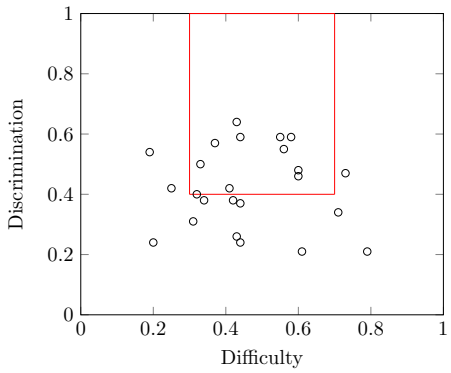

Figure 4.3: Forced-rank Influence

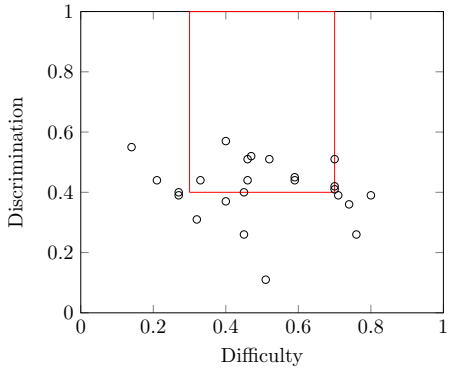

Figure 4.4: Forced-rank Steadiness

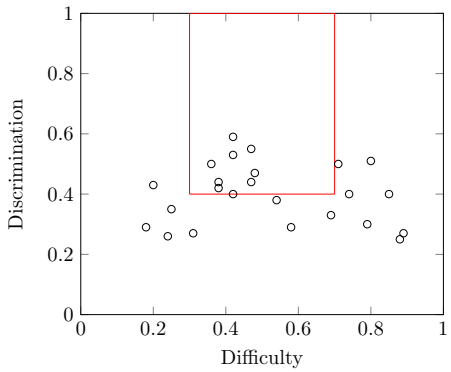

Figure 4.5: Forced-rank Compliance

4.3 Item Analysis Summary

As noted in the previous section, the coverage across the item difficulty or trait continuum is solid. The item discrimination at the various difficulty levels ranges from slightly low to solidly moderate. The authors do note that there is no lower bound on item discrimination of which we are aware. It is more of a common sense argument that higher the levels of discrimination indicate that an item is better at differentiating between individual respondents. It is for this reason that a major focus of the TTI Success Insights continual improvement process is to work towards increased levels of discrimination across the trait continuum.

5 Average Inter-item Correlation

This section presents the results of the average inter-item correlation (IIC)inter item correlation analysis for the TTI Success Insights Style Insights assessment for the English US language. Inter-item correlation measures the assumed linear relationship between the items in a given construct. It is desired that the items correlate strongly enough to be considered to be measuring the same construct, while not being so strong as to introduce redundancy.

Later in this manual a more in depth discussion of reliability in the form of internal consistency is considered, see Section 7 for details. We have also offered a condensed version of the historical development of the so-called α coefficient in Section 3.1. The current discussion centers on the fact that Inter-item correlation and measures of internal consistency are quite closely related. It may be shown that the α coefficient presented in Section 7 is a scaled version of the average inter-item correlation.

For example, 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

$$\alpha_{se} = \frac{K\bar{r}}{(1 + (K-1)\bar{r})}. \tag{5.1}$$

The representation of α_{se} in (5.1) 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 [15],[137],[143],[159],[162], and [178], among many other possibilities.

Of particular note, it is argued in [162] 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 in Section 7, multiple measures are reported. This is consistent with the findings of [162], although the current analyses divert from the details of the information suggested in that article.

5.1 A Brief Description of Average Inter-Item Correlation

Inter-item correlation is a measure of the correlation between the item responses. It is computed using a standard correlation coefficient based on item level responses to the items on 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 on the same 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 [39] 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 6 and 10.

The authors in [39] present a quote of B. F. Green that articulates well the idea we are attempting to describe. See [79] 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." [39], pp. 316.

5.2 Inter Item Correlation Results

The results of the average inter-item correlations are presented in two ways. First, a graphical representation is shown that highlights the boundaries, chosen to be the $0.20 - 0.40$ range, and clearly shows where items fall outside those boundaries, see Figure 5.1. The second presentation is a table of the average inter-item correlations, see Table 5.1. The way to read the table entries is to interpret the row as containing the average inter-item correlation for that item on the appropriate scale, with each column representing one of the four scales.

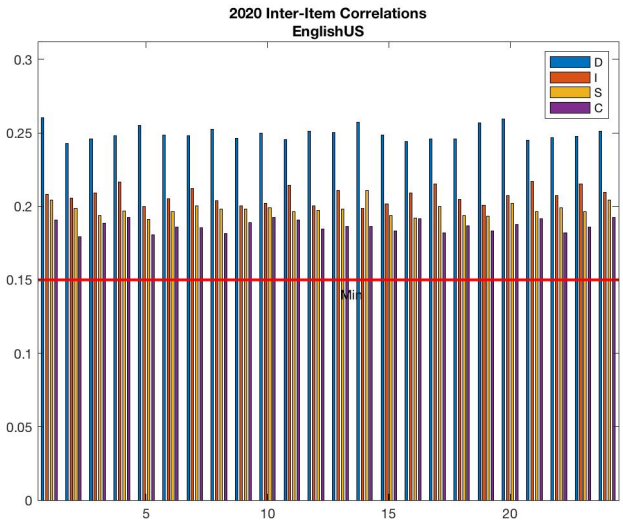

Figure 5.1: English US Style Insights Inter-item Correlation Plot

The vast majority of the items fall well within the desired ranges. The small handful of items that do fall outside acceptable smaller boundaries are still well within the larger range of 0.15-0.50 that some authors state as reasonable. As is noted elsewhere in this manual, no single analysis is used to determine the fit of a single item or group of items. Any items falling outside the stated desired ranges are noted for further analysis in other sections of this manual.

Table 5.1: Average English US Inter-item Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.26	0.21	0.20	0.19
Item 2	0.24	0.21	0.20	0.18
Item 3	0.25	0.21	0.19	0.19
Item 4	0.25	0.22	0.20	0.19
Item 5	0.26	0.20	0.19	0.18
Item 6	0.25	0.21	0.20	0.19
Item 7	0.25	0.21	0.20	0.19
Item 8	0.25	0.20	0.20	0.18
Item 9	0.25	0.20	0.20	0.19
Item 10	0.25	0.20	0.20	0.19
Item 11	0.25	0.21	0.20	0.19
Item 12	0.25	0.20	0.20	0.18
Item 13	0.25	0.21	0.20	0.19
Item 14	0.26	0.20	0.21	0.19
Item 15	0.25	0.20	0.19	0.18
Item 16	0.24	0.21	0.19	0.19
Item 17	0.25	0.22	0.20	0.18
Item 18	0.25	0.20	0.19	0.19
Item 19	0.26	0.20	0.19	0.18
Item 20	0.26	0.21	0.20	0.19
Item 21	0.24	0.22	0.20	0.19
Item 22	0.25	0.21	0.20	0.18
Item 23	0.25	0.22	0.20	0.19
Item 24	0.25	0.21	0.20	0.19
Average	0.25	0.21	0.20	0.19

It should also be noted that we are looking to keep the departure from the mean to a minimum and hence, in some cases, while the score itself is in an acceptable range, its departure away from the true mean may be too large. In these cases, it may be worth considering replacement or potential removal of an item. This determination should not be made based solely on item correlations but is made based on a host of evidence as presented in this report.

5.3 The Attenuation Paradox

The attenuation paradox was first presented in [112] in 1954. The basic idea behind the attenuation paradox is that validity is not a strictly monotonically increasing function of reliability. For completeness, we present the following definition.

Definition 5.1. *A **monotonic** function, $f : A \mapsto B$, is a function between the ordered sets A and B that preserves or reverses the given order.*

For the reader unfamiliar with ordered sets and set theory, a (totally) ordered set is a set X such that for any $x, y \in X$, $x \leq y$, $y \leq x$, or both, in which case $x = y$. A concrete example of an ordered set is the set of all real numbers $\mathbb{R}$. To make things a bit more explicitly clear,

Definition 5.2. *A function f is called monotonically **increasing** (**decreasing**) if for all x, y in the ordered set A such that $x \leq y$ $(x \geq y)$ we have $f(x) \leq f(y)$ $(f(x) \geq f(y))$. If the previous inequality is strict then the function is said to be **strictly** monotonically increasing (decreasing).*

As mentioned in the first paragraph of this section, the attenuation paradox states that validity is not a strictly monotonically increasing function of reliability. In more concrete language, increases in reliability do not lead to a guaranteed increase in validity. In fact, it is quite possible to find oneself in a situation where a push to increase reliability actually leads to a decrease in validity. The following quote is taken from [39], pp. 316.

> Is it not desirable, therefore, to retain highly intercorrelated items in the final scale? No, it is not. This is the essence of the classic attenuation paradox in psychometric theory. Simply put, the paradox is that increasing the internal consistency of a test beyond a certain point will not enhance its construct validity and, in fact, may occur at the expense of validity. One reason for this is that strongly intercorrelated items are highly redundant: Once one of them is included in the scale, the other(s) contribute virtually no incremental information. ... Accordingly, a scale will yield far more information - and, hence, be a more valid measure of a construct - if it contains more differentiated items that are only moderately intercorrelated.

The moral of the story being told here is that lower levels of inter-item correlation tend to enhance validity, especially when considering larger constructs and assuming the items in question are measuring the appropriate construct and the inter-item correlations are at reasonable levels.

6 Corrected Item Total Correlation

In this section we present the (corrected) item-total correlation results for the English US language. 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 the evidence of the internal structure validity of a construct. For a more in-depth discussion of these ideas, one may consult [136] or [156]. 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.

6.1 A Brief Discussion on Corrected Item Total Correlation

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:

$$X = \sum_{i=1}^{n} X_i. \tag{6.1}$$

We may then define the item total correlation of the i^{th} item to be the usual correlation computation between X_i and X:

$$item\ total_i = cor(X_i, X). \tag{6.2}$$

As is seen in (6.1), 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 (6.2).

$$ITC_i = cor(X_i, X - X_i). \tag{6.3}$$

The calculation in (6.3) 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 6.1: 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, [136], pp. 129, presents the information that is summarized in Table 6.1.

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 potentially 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 4.1, partial results of the corrected ITC is presented as part of the item analysis results of that section. Item analysis is a quite encompassing topic given its relative simplicity of implementation. Several items are worth noting here. In Section 5.1, a brief quote attributed to B.F. Green is presented. 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.

6.2 Corrected Item-Total Correlation Results

This section presents the results of the corrected item-total correlation study for the English US language. The results are presented first graphically, in Figure 6.1, with the data presented in Table 6.2.

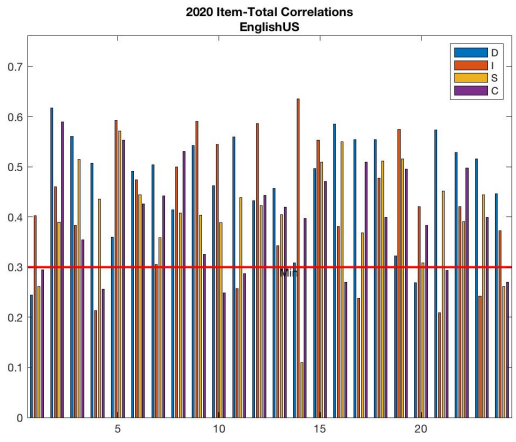

Figure 6.1: English US Style Insights Corrected Item-total Correlation Plot

It is worth discussing multiple scales on this analysis. As is seen in Figure 6.1 several scales have

items at or above the assumed upper bound on the corrected item-total correlation. At the same time, several items are at or near the lower bound shown in the figure.

Table 6.2: English US Corrected Item-total Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.24	0.40	0.26	0.30
Item 2	0.62	0.46	0.39	0.59
Item 3	0.56	0.38	0.51	0.35
Item 4	0.51	0.21	0.44	0.26
Item 5	0.36	0.59	0.57	0.55
Item 6	0.49	0.47	0.44	0.43
Item 7	0.50	0.31	0.36	0.44
Item 8	0.41	0.50	0.41	0.53
Item 9	0.54	0.59	0.40	0.33
Item 10	0.46	0.54	0.39	0.25
Item 11	0.56	0.26	0.44	0.29
Item 12	0.43	0.59	0.42	0.44
Item 13	0.46	0.34	0.40	0.42
Item 14	0.31	0.64	0.11	0.40
Item 15	0.50	0.55	0.51	0.47
Item 16	0.59	0.38	0.55	0.27
Item 17	0.55	0.24	0.37	0.51
Item 18	0.55	0.48	0.51	0.40
Item 19	0.32	0.57	0.52	0.50
Item 20	0.27	0.42	0.31	0.38
Item 21	0.57	0.21	0.45	0.29
Item 22	0.53	0.42	0.39	0.50
Item 23	0.52	0.24	0.44	0.40
Item 24	0.45	0.37	0.26	0.27
Average	0.47	0.42	0.41	0.40
Std. Dev.	0.10	0.13	0.11	0.10

The concepts of small variation in item correlations holds here as well, see Section 5.3. In other words, while the average is at a reasonable level, the variation across the correlation scores for each scale is a bit more than desired. While low corrected item total correlation is likely indicative of a questionable item, further analysis is required to determine the root cause. Any items at or near the lower or upper boundaries are marked for potential further exploration.

6.3 Corrected Item Total Correlation Summary

This section presents the main ideas behind the corrected item total correlation statistic that is used as a measure of how well the individual items are measuring the desired construct. As such, corrected item total correlation may be thought of as a measure of evidence of internal structure validity. While a small handful of items are on the low side, the scales overall perform well with the

average corrected item total correlations in range of providing a moderate amount of information according to the guidelines from the APA as presented in Table 6.1.

7 Internal Consistency as a Measure of Reliability

This section presents the results of internal consistency reliability analyses of the TTI SI Style Insights 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 and coefficient ω with confidence intervals. 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 [73].

The APA makes some significant comments regarding reliability and its relationship to psychometric assessments. First, the APA loosely defines reliability as follows, see [74], chapter 2, pp. 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 it.

The APA also makes the following statement, again in [73], 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, [73], 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 [73], 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.

7.1 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* in addition to time and content sampling. These are the four main areas of reliability the APA chooses to concentrate on in [73]. 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 considered in Section 9 called Generalizability Theory or G-theory for short. The reader interested in more information on G-theory may consult, among many others, [58] for the original work on the topic and [177] for a brief introduction.

Test-retest reliability, also referred to as temporal consistency, measures the reliability with respect to time. In other words, a test-retest approach to reliability measures the responses of the same set of individuals to an assessment taken at two different times and compares the results. See Section 8.1 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 [73], pp. 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 Style Insights assessment.

7.2 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.

Coefficient α was first introduced, for dichotomous items, by Kuder and Richardson, [101], in 1937. A generalized version appears to have first been published in 1941 by Hoyt, see [90], and Jackson and Ferguson, see [92]. Guttman also presented a version of it in 1945, see [82]. The authors should note that the α coefficient has long been attributed to Lee Cronbach, apparently incorrectly so. The authors in [38] and [86] present a nice history of how the α coefficient developed and why it is attributed to Cronbach. In particular, the following quote is from [86]:

> A few years after Guttman's introduction, alpha was popularized by Cronbach, [56]. The term alpha came from Cronbach, who expected that alpha would just be the first of a range of similar measures, which could be given subsequent letters from the Greek alphabet.

It appears that the main reason for the association between Cronbach and the coefficient α is that he was the first to give it a name. To be sure, Cronbach has been a big contributor to the literature in this area over the years. As such, the authors of this manual do not intend to be critical of Cronbach. We merely seek to get to the truth of the matter.

Since that time, 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 5 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 [15], [137] ,[143] ,[159], [162], and [178], 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 [57]. The authors mentioned in [15], [137] ,[143] ,[159], [162], and [178] offer alternatives, among which are reporting α coefficient with confidence intervals as presented in Section 7.2.1. Additionally, coefficient ω, with confidence intervals is also presented, see Sections 7.3 and 7.3.1.

7.2.1 English US α Coefficient Results

This section presents the results of the internal consistency study for the English US language. The data are presented in Table 7.1.

Table 7.1: Average English US α Coefficient Data
TTI SI Style Insights

Est./Scale	D	I	S	C
α	0.890	0.865	0.855	0.850
S.E.	0.001	0.001	0.001	0.001
Lower CI	0.889	0.864	0.854	0.849
Upper CI	0.891	0.866	0.856	0.851

For interpreting reliability coefficients in general, Table 7.2 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 [59]. However, on page 184 of [10], 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 7.2: Common Reliability
Coefficient Interpretation, see [59]

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

The author in [10] 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, [131] states that 0.70 is acceptable during development, 0.80 for "basic research scales". and 0.90 for clinical settings are appropriate minimums. [142] recommends a minimum of 0.80 without the

same restrictions as [131]. As a final example, [97] suggests values of 0.95 for clinical settings. [10] 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 7.2 or the merits of any particular bounds that have been discussed over the years. We present Table 7.2 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 [59] states unequivocally that these bounds are his preference and that "I cannot defend them on strictly rational grounds."

As discussed in Section 3, the lack of consensus and, at times, guidance from the academic community on subjects such as validity and reliability of psychometric assessments places a potential barrier in front of practitioners. Organizations such as the APA or National Council on Measurement in Education, the Mental Measurement Yearbook, etc., have great influence in these areas. Yet, even in [74], there is little to no guidance as to what constitutes an acceptable value of an α coefficient. It is then left to the judgment of the developer, using the available tools, to determine whether a scale shows an acceptable level of evidence of internal consistency reliability.

Combining the information in Table 7.2 with the results presented in Table 7.1, it is seen that the TTI SI EQ assessment scales all show solid evidence of internal consistency as measured by the α coefficient. Additionally, the standard errors (S.E. in Table 7.1) are all relatively small with respect to the magnitude of the coefficients leading to relatively tight confidence intervals. One is then led to conclude that, based on the α coefficient analysis, that all 5 scales of the TTI Success Insights Style Insights assessment score in the upper end of the Respectable range to the Very Good range.

7.3 A Brief Discussion of Coefficient ω

The ω coefficient comes from the work of Roderick P. McDonald, see for example [126]. Unlike the α coefficient which is a scaling of inter-item correlation, ω is based on a factor analytic approach. For a simplistic view of the ω calculation, one assumes a single factor model fit of the assessment data. If the fit shows a homogeneous assessment, compute the ω coefficient by squaring the sum of the factor loading parameters and divide by the total variance of the assessment scores.

As noted in Section 7.2.1, a push has been made over the last couple decades to move from internal consistency estimates such as the α coefficient to more robust measures such as ω, see [61] or a[137], to name a few. ω computation takes the estimation of internal consistency coefficients into the realm of factor analytic methods, in some sense bridging part of the gap between reliability and validity. For example, one estimate of construct validity may be obtained by considering the variance explained as reported by a factor analytic approach to analyzing assessment data.

7.3.1 English US ω Coefficient Results

This section presents the results of the internal consistency study for the English US language based on the McDonald's ω coefficient.

Table 7.3: Average English US ω Coefficient Data
TTI SI Emotional Quotient

Est./Scale	D	I	S	C
ω	0.898	0.866	0.866	0.855
S.E.	0.005	0.006	0.006	0.006
Lower CI	0.887	0.853	0.855	0.843
Upper CI	0.905	0.878	0.878	0.867

Combining the information in Table 7.2 with the results presented in Table 7.3, it is seen that the TTI Success Insights Style Insights assessment scales all show solid evidence of internal consistency as measured by the ω coefficient. The scores range from a low of 0.855 for Compliance to a high of 0.898 for Dominance. The standard errors (S.E. in Table 7.3) are all relatively small with respect to the magnitude of the coefficients leading to relatively tight confidence intervals. It is again the case that the scales of the TTI Success Insights Style Insights assessment show reasonable evidence of internal consistency based on the ω coefficient analysis.

7.4 α-if-Item-Deleted Analysis

This section presents the so-called α-if-item-deleted analysis. The main idea in this section is to compute the α coefficient for the scale with each item deleted from consideration. If abnormal changes in the value of the α coefficient are noted, the item should be flagged for further consideration in other areas of analyses being conducted. It may be the case that a given item is not a good fit with the remainder of the items used for measuring the scale. As is noted consistently in this technical manual, a decision to remove or replace an item is not made based on any single statistical measure. Rather, decisions are made based on the totality of information available.

Recall that the α coefficients for each of the six scales are presented in Table 7.1 above. In the following tables, the value of the α column represents the α coefficient for the scale with the respective item removed. This value should be considered in relation to both the original α coefficient in Table 7.1 as well as the remaining α if item deleted values in the respective tables. We begin with the Dominance scale.

Table 7.4: English US α-if-Item-Deleted Data
Dominance, $\alpha = 0.889$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.890	Item 7	0.885	Item 13	0.886	Item 19	0.889
Item 2	0.882	Item 8	0.887	Item 14	0.889	Item 20	0.891
Item 3	0.884	Item 9	0.884	Item 15	0.885	Item 21	0.883
Item 4	0.885	Item 10	0.886	Item 16	0.883	Item 22	0.884
Item 5	0.889	Item 11	0.884	Item 17	0.883	Item 23	0.884
Item 6	0.885	Item 12	0.887	Item 18	0.883	Item 24	0.886

Table 7.5: English US α-if-Item-Deleted Data
Influence, $\alpha = 0.865$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.861	Item 7	0.864	Item 13	0.863	Item 19	0.855
Item 2	0.859	Item 8	0.858	Item 14	0.853	Item 20	0.860
Item 3	0.862	Item 9	0.855	Item 15	0.856	Item 21	0.866
Item 4	0.867	Item 10	0.857	Item 16	0.862	Item 22	0.860
Item 5	0.855	Item 11	0.865	Item 17	0.865	Item 23	0.866
Item 6	0.859	Item 12	0.855	Item 18	0.859	Item 24	0.862

Table 7.6: English US α-if-Item-Deleted Data
Steadiness, $\alpha = 0.855$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.855	Item 7	0.852	Item 13	0.850	Item 19	0.846
Item 2	0.850	Item 8	0.850	Item 14	0.859	Item 20	0.853
Item 3	0.846	Item 9	0.850	Item 15	0.846	Item 21	0.848
Item 4	0.849	Item 10	0.851	Item 16	0.847	Item 22	0.851
Item 5	0.844	Item 11	0.849	Item 17	0.851	Item 23	0.849
Item 6	0.849	Item 12	0.849	Item 18	0.846	Item 24	0.856

Table 7.7: English US α-if-Item-Deleted Data
Compliance, $\alpha = 0.849$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.848	Item 7	0.843	Item 13	0.844	Item 19	0.841
Item 2	0.837	Item 8	0.839	Item 14	0.845	Item 20	0.845
Item 3	0.846	Item 9	0.847	Item 15	0.842	Item 21	0.848
Item 4	0.849	Item 10	0.849	Item 16	0.848	Item 22	0.841
Item 5	0.839	Item 11	0.848	Item 17	0.841	Item 23	0.845
Item 6	0.844	Item 12	0.843	Item 18	0.845	Item 24	0.849

7.5 A Note on Comparing Reliability and Validity

As noted in Section 5.3, having high levels of internal consistency measures may be good, but can also be indicative of overly homogeneous scales. A common mistake in the field of assessments is to (inadvertently) ignore measures of validity and focus on measures of internal consistency. There are many reasons for this approach. First and foremost is the lack of overall consensus as to what actually constitutes "validity". Further, there is not a complete agreement as to what are the different types of validity. As an example, Messick aligns with 6 different areas of validity to consider, see [128] and [129], while the most recent APA guidelines suggest 5 areas, see [156].

Different authors combine different aspects of other authors' opinions on validity measures in different ways. At the end of the day, the concept of what constitutes validity or adequate levels of

evidence of validity seems confusing. On the other hand, computing the α coefficient to measure evidence of internal consistency is straight forward. Even if we move to the more complex measures of internal consistency such as ω, black-box software packages exist for the average person to pass some data through and get a result. Tables such as that shown Table 7.2 exist to lend immediate interpretation. It is tempting, therefore, to put the best foot forward in an area that is easy to understand for the assessment developer as well as the end user of any product based on the underlying assessment.

Then, as discussed in Section 5.3 and [112], the Attenuation Paradox comes into play. Unfortunately, this is a counterintuitive concept, abstract and even esoteric to some, that tells the assessment developer that it may actually be necessary to lower measures of internal consistency in order to obtain acceptable levels of evidence of validity. The concept of lowering internal consistency measures to improve evidence of validity measures is more related to α than ω based on the fact that α is more directly related to inter-item correlations.

For example, one may improve internal consistency measures by increasing the inter-item correlations or by increasing the number of items assessed. In either case, the realization is to increase the homogeneity of the assessment, thereby decreasing the likelihood that the construct domain is any more adequately covered at the higher level of measure of internal consistency. The end result, therefore, is that increasing homogeneity (internal consistency) decreases the coverage of the content domain (validity).

It is not that the conclusion is that it is necessary to decrease internal consistency to obtain validity. Rather, it is the conclusion that to adequately cover a content domain it may be necessary to construct a less homogeneous assessment. The end result is that while internal consistency measures may decrease, validity measures increase. Reliability is an important measure, but validity is more so. If necessary, sacrificing measures of reliability to obtain validity develops a better overall assessment, at least according to the currently accepted standards of the APA and authors such as [39, 112], among many others.

8 Temporal Consistency as a Measure of Reliability

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. We present the APA's view of test-retest reliability, followed by rest-retest study results.

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

Temporal consistency is one of the four main areas of reliability endorsed by the APA, see chapter two of [74]. According to the author of this chapter, [73], 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 this definition of stability.

According to [73], pp. 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 multiple applications of the same assessment over time. What is truly important is that the relative standing among 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 [73] 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 that may differentially impact a respondent's scores with respect to behavior. This assessment is measuring behavior styles. While it is possible for the behavior an individual to change over time, we do not expect changes over relatively short time periods. Waiting for an exceedingly long time between administrations may cause an underestimate of the reliability of the assessment.

The only guidance the APA provides on the subject is as follows, see [73], 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 Style Insights 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 Style Insights assessment is 24 frames. These frames are split along four scales and the respondent is asked to rate each statement in relation to the remaining statements in terms of how much the statement is a reflection of the individual's immediate reaction to the statements. 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 24 frames on the TTI Success Insights assessment are currently presented as a four-item forced-rank frame. We again note that the length of time between and the length of the assessment discourage memorization. We do note that while it is possible to memorize the constructs and their corresponding statements, we feel it is unlikely that a large enough proportion of any test-retest population has the desire to do so.

Finally, the nature of the characteristics being measured is addressed as follows. This assessment is measuring behavior. It is possible that memorization becomes a problem, but TTI Success Insights feels this is unlikely. Given that this assessment is measuring behavior, it is more likely that the state of emotions of the individual at the time of administration may directly impact the reliability of the assessment and we address such issues in Section **??**. Memorization issues are unlikely to be a major problem for TTI Success Insights Style Insights assessment.

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 [73], 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 Style Insights) 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.

8.2 TTI Success Insights 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 continual improvement process. For a sample of previous test-retest studies, see [24].

The current test-retest iteration is comprised of a subset of the overall TTI Success Insights Style Insights database. The data spans the time frame January 1, 2015 to March 25, 2020. The limited

demographic information available at the time of publication of this manual is presented in Table 8.1. As is seen, there is approximately a 50/50 split on the male/female breakdown in the population of slightly more than 33,000. The mean time between administrations of the test is slightly more than 16 months, with the median time between administrations just more than 14 months.

Table 8.1: Test-Retest Data Breakdown
TTI Success Insights Style Insights

	1/1/2015 to 3/25/2020
Elapsed Time	3-38 Months
Mean Elapsed Time	486 Days
Median Elapsed Time	420 Days
Population	33,315
Males	16,825 (50%)
Females	16,440 (50%)

In many, although not all, cases the TTI Success Insights Style Insights assessment is utilized as part of a training and/or coaching program. As such, there is an expectation of "improvement" in scoring that will potentially degrade the overall test-retest correlation results, especially relevant to smaller data samples. By improvement it is meant that during a coaching or training process, aspects of individuals' behavior, motivation, or other constructs being measured may be identified as areas on which to work for a given reason, and improvement is meant to reflect a movement towards a specific goal as measured by the variables underlying the Style Insights assessment. As noted in an earlier section, specifically quoting [73], measures of temporal reliability are measures of "persistent relative standing within the group rather than consistent absolute value of score".

Table 8.2: Test-Retest Correlations 01/01/2015 to 03/25/2020
TTI Success Insights Style Insights

Scale	D	I	S	C
D	0.80			
I		0.84		
S			0.79	
C				0.82

As such, it is to be expected that there may be some negative impact on the relative standing of individuals when comparing those being coached and trained in improving the specifics of individual as compared to those who may not be participating in such a program. At the time of writing this manual, there is no distinguishing database element to determine whether an individual participated in a coaching or training program, or whether an individual took the assessment for other purposes.

Table 8.2 contains the correlation results between the first and second administrations of the TTI Success Insights Style Insights Insights assessment in the data from January 1, 2015 to March 25, 2020. The results are solidly in the 0.79 to 0.84 range with all correlations computed with statistical significance at less than a 0.001 level. Based on the results of the correlation analysis, we conclude that the TTI Success Insights Style Insights temporal consistency is clearly at an acceptable level.

8.3 TTI Success Insights 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 [73]. The results of the most recent version of the test-retest analysis, presented in Table 8.2, show solid evidence of test-retest reliability for the time period from January 1, 2015 to March 25, 2020.

The next section presents a slightly different view of the test-retest or temporal consistency analysis. Both data sets used in this section for a correlation analysis are subjected to the approach of Cronbach, et. al., in [58] known as Generalizability or G-theory.

9 Generalizability Theory Applied to Test-retest Data

As noted in Section 7.5, reliability is one of the most important criteria against which assessment scores are evaluated. There are many publications addressing the shortcomings of measures of internal consistency, such as the α coefficient (as noted in Section 7.2, incorrectly attributed to Cronbach, [38]), see, e.g., [15], [55], [61], [64], [137], [143], [159], or [178]. Even Cronbach was quite critical of the α coefficient towards the end of his career, see [57].

As an example of the shortcomings of some of the measures of internal consistency, it is widely known that most are functions, generally speaking, of two main components. The first is inter-item correlation and the second is the number of items on the assessment or scale. In essence, one may increase the reliability of an assessment or scale by increasing the corresponding number of items (assuming they measure the same constructs as the assessment or scale) or by increasing the inter-item correlation among the items by choosing a more homogeneous set of items.

These approaches point to the main issue of the attenuation paradox. Both approaches may be used to increase the reliability of an assessment or scale. However, neither approach addresses the potential lost validity by increasing the number of items while maintaining or increasing the level of homogeneity of the items generating the scales. In other words, by following this approach, one is strictly attempting to increase reliability without taking into account the possible, even likely, decrease in validity of the assessment or scales of the assessment.

There appear to be (at least) two main goals at this point. The first is to have a solid measure of the evidence of the reliability of an assessment without compromising the assessment's validity. The second is to attempt to uncover the sources of variation in scoring and attempt to assign the proportion of the variability of assessment scores that is due to each source of variation. This second goal is, in the opinion of the authors of this manual, a beginning of the bridging of the gap between reliability and validity. In other words, the more sources of variation that can be discovered with as much variation as possible being attributed to the items, the more valid an assessment should be. This last statement assumes that the proportion of variation accounted for by a given source of variation is appropriate to the intended study.

Generalizability theory (G-theory) attempts to answer these questions in a theoretically sound way, see [28], [58], [152], or [177]. The remainder of this section provides a brief introduction to the concepts of generalizability and how it may be used to measure test-retest reliability. We present both a one- and two-facet model for instructional purposes, although only the one-facet model is applied to the assessment data. The data considered are the same data from the earlier section on test-retest reliability, see Section 8.

9.1 The Basics of One- and Two-Facet Generalizability Theory Models

This section presents the basics of one- and two-facet Generalizability theory (G-theory) models mostly by example. The examples are taken from an article by Marcoulides, see [119], the APA testing and assessment handbook, [177], and the original work on the subject, [58].

Classical test theory assumes that an individual's observed score is comprised of two components, the true score that represents stable, non-random individual differences, and an error term:

$$X_{pi} = T_{pi} + e_{pi}, \tag{9.1}$$

where X_{pi} is the individual's observed score on item i, T_{pi} is the true score, and e_{pi} is the measurement error (not truly random). On the other hand, a G-theory approach with one facet, in this

case the items, has the following formulation of an individual's score on an item:

$$X_{pi} = \mu + \mu_p^* + \mu_i^* + \mu_{pi}^* + e_{pi}, \tag{9.2}$$

where μ is the overall mean in the population of both individuals and items, μ_p^* is the score effect attributable to person p, μ_i^* is the score effect attributable to item i, μ_{pi}^* is the score effect attributable to the interaction of person p with item i, and e_{pi} is the unaccounted for error.

It is noted in both [119] and [58] that there is little to delineate the difference between the effects of μ_{pi}^* and e_{pi}. With this in mind, it is common to allow for a single term to account for both, originally introduced with the notation $\mu_{pi,e}^*$ in [58]. In this notation, (9.2) may be written as

$$X_{pi} = \mu + \mu_p^* + \mu_i^* + \mu_{pi,e}^*. \tag{9.3}$$

In order to facilitate estimation, it is convenient to rewrite (9.3) as

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu), \tag{9.4}$$

where $(\mu_p - \mu)$ represents the person effect μ_p^*, $(\mu_i - \mu)$ represents the item effect μ_i^*, and $(X_{pi} - \mu_p - \mu_i + \mu)$ represents the residual effect $\mu_{pi,e}^*$. With this notation, it is possible to estimate the quantities of interest μ, μ_p, and μ_i. For example, one may estimate μ_p by taking the expectation of the X_{pi} over all persons p, which is different than taking the same expectation over all items i.

Keeping in mind that the end goal is to give an estimate of reliability, and that reliability is the ratio of certain variances, one may now take the information from the preceding paragraph and develop an estimate of variances based on many possible approaches. In practice, the most popular method the authors of this manual are aware of is to use the mean squares from an analysis of variance (ANOVA) and equate these values to their theoretically expected values. In other words, we know the values necessary to compute the ANOVA which generates estimates of the mean square errors. Further, we have the theoretical values of the mean square errors in terms of variances. We desire estimates of these variances in order to estimate reliability coefficients.

For the current example we have the following theoretical relationships as presented in [58], pp. 44.

$$\begin{array}{lcl} EMS_p & = & \sigma_{pi,e}^2 + n_i\sigma_p^2 \\ EMS_i & = & \sigma_{pi,e}^2 + n_p\sigma_i^2 \\ EMS_{pi,e} & = & \sigma_{pi,e}^2 \end{array} \tag{9.5}$$

Note that the left hand side of (9.5) is completely known from the ANOVA calculations and the right hand side is a triangular system that may be easily solved by substitution methods. Of course, for systems with many facets it is much easier to use the power of modern computing systems using matrix algebra. The elements σ_p^2, σ_i^2, and $\sigma_{pi,e}^2$ represent the unknown variances we are interested in computing, and the values n_p and n_i represent the number of individual respondents and the number of items on the assessment respectively.

Now that we have estimates of the variances for the relevant pieces of the puzzle, we are interested in what is called the δ-type error in the literature. The δ-type error is of primary concern when researchers are interested in decisions that rank order individuals, such as the test-retest data from Section 8. All sources of variation that involve individuals are removed from consideration and we compute the following:

$$\sigma_\delta^2 = \frac{\sigma_{pi,e}^2}{n_i}. \tag{9.6}$$

σ_δ^2 may be interpreted as the variance of the errors for relative decisions. We are now in a position to define the generalizability coefficient in the sense that we may compute the ratio of the universal score variance to the expected observed score variance, see [177].

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}. \tag{9.7}$$

(9.7) is analogous to the usual standard reliability coefficients such as α and ω as presented in Section 7. In contrast to the relative error is the absolute error which is concerned when both rank ordering and differences in averages may be relevant. In the one facet case, we are interested in both the residual effect and the effect due to items. Hence, we compute the following:

$$\sigma_\Delta^2 = \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pi,e}^2}{n_i}. \tag{9.8}$$

A final piece comes into play through the index of dependability, defined as

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}. \tag{9.9}$$

At this point, it is convenient to present a hypothetical example to help solidify the ideas presented so far in this section. This example and data in Table 9.1 are pulled from [119]. The details of the computations of the values for EMS_p, EMS_i, and $EMS_{pi,e}$ are lengthy and are based on standard ANOVA calculations. The data in Table 9.1 is a hypothetical example of five individuals with scores on five distinct items.

Table 9.1: Hypothetical Example of 5 Individuals and a Single Facet

	Items				
Person	**1**	**2**	**3**	**4**	**5**
1	9	6	6	8	8
2	8	5	6	6	6
3	9	8	7	8	8
4	7	5	6	4	4
5	7	5	6	3	4

Based on the data provided in Table 9.1, the following values are computed.

$$\begin{aligned} EMS_{pi,e} &= 0.84 \\ EMS_i &= 4.34 \\ EMS_p &= 8.74 \end{aligned} \tag{9.10}$$

The results in (9.10), substituted into (9.5), lead to the following values for the variances of interest.

$$\begin{aligned} \sigma_{pi,e}^2 &= 0.84 \\ \sigma_i^2 &= \frac{EMS_i - EMS_{pi,e}}{n_p} = \frac{4.34-0.84}{5} = 0.70 \\ \sigma_p^2 &= \frac{EMS_p - EMS_{pi,e}}{n_i} = \frac{8.74-0.84}{5} = 1.58 \end{aligned} \tag{9.11}$$

Continuing, we compute the σ_δ^2 and σ_Δ^2 values, using (9.6) and (9.8), respectively.

$$\begin{array}{rcccl} \sigma_\delta^2 & = & \frac{\sigma_{pi,e}^2}{n_i} & = & \frac{0.84}{5} = 0.17 \\ \sigma_\Delta^2 & = & \frac{\sigma_i^2+\sigma_{pi,e}^2}{n_i} & = & \frac{0.70}{5}+\frac{0.84}{5} = 0.31 \end{array} \tag{9.12}$$

Finally, we may compute the generalizability coefficient and dependability index using (9.7) and (9.9), respectively.

$$\begin{array}{rcccl} E\rho_\delta^2 & = & \frac{\sigma_p^2}{\sigma_p^2+\sigma_\delta^2} & = & \frac{1.58}{1.58+0.17} = 0.90 \\ \Phi & = & \frac{\sigma_p^2}{\sigma_p^2+\sigma_\Delta^2} & = & \frac{1.58}{1.58+0.31} = 0.31 \end{array} \tag{9.13}$$

We now wish to extend these ideas to a situation in which we have more than one facet. The details of an n-facet model are beyond the scope of this manual and the interested reader is referred to any of [28], [58], or [177]. We present, again by example, a two-facet model and briefly discuss the ideas behind extending the two-facet model to a three-facet model.

The following example is presented in [177] with a shortened, similar example provided in [119]. Consider a group of 20 student teachers preparing lesson plans that are to be graded by two different raters. There are a total of four exercises for each of the 20 individuals and two raters to provide scoring for each of the four exercises for each of the 20 individuals. A schematic style view of the main idea behind the problem at hand is presented in Table 9.2.

Table 9.2: Schematic View of Teacher Lesson Plan Example

Teacher	**Rater 1**				**Rater 2**			
Lesson Plan	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 1**	**Task 2**	**Task 3**	**Task 4**
1	$X_{1,1,1}$	$X_{1,2,1}$	$X_{1,3,1}$	$X_{1,4,1}$	$X_{1,1,2}$	$X_{1,2,2}$	$X_{1,3,2}$	$X_{1,4,2}$
2	$X_{2,1,1}$	$X_{2,2,1}$	$X_{2,3,1}$	$X_{2,4,1}$	$X_{2,1,2}$	$X_{2,2,2}$	$X_{2,3,2}$	$X_{2,4,2}$
3	$X_{3,1,1}$	$X_{3,2,1}$	$X_{3,3,1}$	$X_{3,4,1}$	$X_{3,1,2}$	$X_{3,2,2}$	$X_{3,3,2}$	$X_{3,4,2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	$X_{20,1,1}$	$X_{20,2,1}$	$X_{20,3,1}$	$X_{20,4,1}$	$X_{20,1,2}$	$X_{20,2,2}$	$X_{20,3,2}$	$X_{20,4,2}$

In this case we now have two facets, namely the raters and the tasks with which each of the individuals must interact. This implies that we must consider sources of variation in addition to the individual and the item. We must extend the ideas of equation (9.4) to account for the additional complexity. Equation (9.14) shows the "easy" formula as presented in [177], as well as in the original source [58].

It is noted that the values μ, μ_α, and $\mu_{\alpha\beta}$ are all theoretically computed using an expected value. For example, μ_p is computed as the expectation of X_{ptr} with respect to t and r, $\mu_p = E_t E_r X_{ptr}$. Similarly, $\mu_{pt} = E_r X_{ptr}$. The remainder of the values are computed in a similar manner replacing the indices in the fashion just described.

We now find ourselves in the situation of having seven separate mean values that represent potential effects that are introduced during the entire assessment process. In this case, the assessment process includes the scoring as generated by the raters, not just the taking of the assessment (performing the tasks).

$$
\begin{aligned}
X_{ptr} &= \mu && \text{(grand mean)} \\
&\quad +(\mu_p - \mu) && \text{(person effect)} \\
&\quad +(\mu_t - \mu) && \text{(task effect)} \\
&\quad +(\mu_r - \mu) && \text{(rater effect)} \\
&\quad +(\mu_{pt} - \mu_p - \mu_t + \mu) && \text{(person} \times \text{task effect)} \\
&\quad +(\mu_{pr} - \mu_p - \mu_r + \mu) && \text{(person} \times \text{rater effect)} \\
&\quad +(\mu_{tr} - \mu_t - \mu_r + \mu) && \text{(task} \times \text{rater effect)} \\
&\quad +(X_{ptr} - \mu_{pt} - \mu_{pr} - \mu_{tr} + \mu_p + \mu_t + \mu_r - \mu) && \text{(residual effect)}
\end{aligned}
\tag{9.14}
$$

These seven items are now relabeled by their common names from the ANOVA procedure similar to that shown in (9.5) and defined by their theoretical values as presented in both [58] and [177].

$$
\begin{aligned}
EMS_p &= \sigma^2(ptr,e) + n_r\sigma^2_{pt} + n_t\sigma^2_{pr} + n_t n_r \sigma^2_p \\
EMS_t &= \sigma^2(ptr,e) + n_p\sigma^2_{tr} + n_r\sigma^2_{pt} + n_p n_r \sigma^2_t \\
EMS_r &= \sigma^2(ptr,e) + n_p\sigma^2_{tr} + n_t\sigma^2_{pr} + n_p n_t \sigma^2_r \\
EMS_{pt} &= \sigma^2(ptr,e) + n_r\sigma^2_{pt} \\
EMS_{pr} &= \sigma^2(ptr,e) + n_t\sigma^2_{pr} \\
EMS_{tr} &= \sigma^2(ptr,e) + n_p\sigma^2_{tr} \\
EMS_{ptr,e} &= \sigma^2(ptr,e)
\end{aligned}
\tag{9.15}
$$

It is noted that (9.15) again defines an almost triangular system that is easily solved by using a simple substitution method with the values on the left hand side of the equation known a priori from the ANOVA procedure. We provide the analogous computations for the values of σ^2_δ and σ^2_Δ next.

$$
\begin{aligned}
\sigma^2_\delta &= \frac{\sigma^2_{pt}}{n_t} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{ptr,e}}{n_t n_r} \\
\sigma^2_\Delta &= \frac{\sigma^2_t}{n_t} + \frac{\sigma^2_r}{n_r} + \frac{\sigma^2_{pt}}{n_t} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{tr}}{n_t n_r} + \frac{\sigma^2_{ptr,e}}{n_t n_r}
\end{aligned}
\tag{9.16}
$$

Finally, the generalizability coefficient and dependability index are defined in the same way as previously, see (9.7) and (9.9), reproduced for convenience here.

$$
\begin{aligned}
E\rho^2 &= \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta} \\
\Phi &= \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta}
\end{aligned}
\tag{9.17}
$$

The extension to higher facet models should be fairly obvious at this stage, although the details may seem a bit overwhelming. However, if one notes the symmetry that is developing when considering the defining equations for the one- and two-facet models, this complexity abates somewhat. As an example, consider a three-facet model of the form $(i \times j \times k \times p)$, to use the notation of [58], and note that we have made the simplifying assumption that our needs are fulfilled by a model of this type

rather than the multitude of other possible three-facet models. Then consider the representative EMS-equations for this model as presented in (9.18).

$$\begin{aligned} EMS_{p} &= \sigma^2(pijk,e) + n_i\sigma^2_{pjk} + n_j\sigma^2_{pik} + n_k\sigma^2_{pij} + n_in_j\sigma^2_{pk} \\ &\quad + n_in_k\sigma^2_{pj} + n_jn_k\sigma^2_{pi} + n_in_jn_k\sigma^2_{p} \\ EMS_{pi} &= \sigma^2(pijk,e) + n_j\sigma^2_{pik} + n_k\sigma^2_{pij} + n_jn_k\sigma^2_{pi} \\ EMS_{pij} &= \sigma^2(pijk,e) + n_k\sigma^2_{pij} \\ EMS_{ptr,e} &= \sigma^2(ptr,e) \end{aligned} \tag{9.18}$$

The remainder of the coefficients may be filled in using symmetry and other properties that are apparently related to a particular set of coefficients related to the binomial theorem, a statement made without an offer of proof at this time. As an example, it is hypothesized that there are a total of 15 defining equations for the three-facet model which is the sum of the coefficients of the binomial $(a+b)^4$ minus one. This equation holds for one and two facet models as well, but we only assume the forms $(i \times j)$, $(i \times j \times k)$, and $(i \times j \times k \times p)$. The authors make no claims on other forms of the facet models that are also included in generalizability theory. It is likely that under certain assumptions one may be able to formulate and prove a proposition via mathematical induction on the number of facets.

As a final presentation of this portion of the section on G-theory, we formulate the test-retest problem as a two-facet model. The schematic of such a formulation is presented in Table 9.3.

Table 9.3: Style Insights Scale X Test-Retest as a Two-Facet G-Theory Model

Individual	**Time 1**				**Time 2**			
Respondent	**Item 1**	**Item 2**	$\cdots$	**Item m**	**Item 1**	**Item 2**	$\cdots$	**Item m**
1	$X_{1,1,1}$	$X_{1,2,1}$	$\cdots$	$X_{1,m,1}$	$X_{1,1,2}$	$X_{1,2,2}$	$\cdots$	$X_{1,m,2}$
2	$X_{2,1,1}$	$X_{2,2,1}$	$\cdots$	$X_{2,m,1}$	$X_{2,1,2}$	$X_{2,2,2}$	$\cdots$	$X_{2,m,2}$
3	$X_{3,1,1}$	$X_{3,2,1}$	$\cdots$	$X_{3,m,1}$	$X_{3,1,2}$	$X_{3,2,2}$	$\cdots$	$X_{3,m,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$X_{n,1,1}$	$X_{n,2,1}$	$\cdots$	$X_{n,m,1}$	$X_{n,1,2}$	$X_{n,2,2}$	$\cdots$	$X_{n,m,2}$

The table shows a generic scale of m items. As is discussed in Section 2, the TTI Success Insights Style Insights assessment has four scales, each of which has 24 items. The index n represents the number of respondents in the data set under consideration and the $Time$ 1 and $Time$ 2 labels for the final facet represent the two separate administrations of the assessment.

9.2 One-Facet G-theory Model Applied to TTI Success Insights Style Insights Test-Retest Data

This section extends the ideas of the usual test-retest correlation study to a one facet model using generalizability theory. The idea here is to treat the "time" variable as a source of variance, in addition to the variance present in the individual respondents. Given that the respondents' scores

are the item of interest, they are not included as a facet, hence the use of a one facet model, the only facet being two separate administrations of the assessment.

Table 9.4: Hypothetical Example of n Individuals and One Facet ("Time")

	D Scores	
Person	**Time 1**	**Time 2**
1	31	38
2	32	35
3	31	38
4	37	30
5	37	33
⋮	⋮	⋮
n	41	41

Table 9.4 shows how the data for a test-retest sample may be formulated into the one-facet modeling approach, see for example Table 9.1 in Section 9.1. The way to interpret the data in Table 9.4 is that each row represents an individual respondent and each column the respondent's score at Time 1 or Time 2. As an example, the final row, row n, shows that the last respondent listed in the data set scores 41 on the D scale on both administrations of the assessment. We then compute the variance table as shown in Table 9.5.

9.2.1 One-Facet Model Results for 1/1/2015 to 3/25/2020

This section presents the results of the G-theory based on a one-facet model for data in the time period 1/1/2015 to 3/25/2020.

Table 9.5: Style Insights Dominance Scale Test-Retest
One-Facet G-Theory Model 1/1/2015 to 3/25/2020

Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**Expected Mean Square**	**Component Variance**	**Component % of Variance**
Persons (p)	33,314	$1.2e^7$	352	$\sigma^2_{pt,e} + n_t\sigma^2_p$	156	79.7
Time (t)	1	7192	7192	$\sigma^2_{pt,e} + n_p\sigma^2_t$	0.21	0.001
Residual (pt,e)	33,314	$1.32e^6$	40	$\sigma^2_{pt,e}$	40	20.1

The first thing of note in Table 9.5 is the very small, positive component variance for the Time facet, especially when compared to the Persons source of variability. This low component variance, accounting for well under 1% of the total variance in this model, indicates that Time is not a major source of variability for the Dominance scale. In this setting we are interested in determining precisely whether Time is a major source of variation in the scoring of the scales of the TTI Success Insights Style Insights assessment. It does not appear that this is the case for the Dominance scale.

The second main piece of information to take away from this analysis is that nearly 80% of the variability in the model under consideration is accounted for by the Persons source of variation.

This does imply that approximately 20% of the variation is unaccounted for. The combination of the preceding information may be interpreted to show that more than 50% of the variation is accounted for by Persons, yet a portion of the variation is not accounted for and further analysis may be warranted.

Table 9.6 shows the results for the Influence scale. In this case we see that in excess of 83% of the component variance is explained by the Persons component. Influence does show the largest component variance due to time, yet this scale still only accounts for less than 1% of the total component variance. This appears to be quite solid performance with less than 17% of the variance not accounted for, as well as the majority of the variance accounted for attributed to the Persons facet.

Table 9.6: Style Insights Influence Scale Test-Retest
One-Facet G-Theory Model 1/1/2015 to 3/25/2020

Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**Expected Mean Square**	**Component Variance**	**Component % of Variance**
Persons (p)	33,314	$9.7e^6$	291	$\sigma^2_{pt,e} + n_t\sigma^2_p$	132	83.5
Time (t)	1	22.67	22.67	$\sigma^2_{pt,e} + n_p\sigma^2_t$	0.00	0.004
Residual (pt,e)	33,314	$8.7e^5$	26	$\sigma^2_{pt,e}$	26.11	16.5

The Steadiness scale is presented in Table 9.7, again showing quite similar results to the Dominance scale. In excess of 78% of the component variance is explained by the Persons component and a small component variance for the Time facet. The Steadiness scale shows a similar percentage of unaccounted for variance with an even lower percentage of variance attributed to the Time facet.

Table 9.7: Style Insights Steadiness Scale Test-Retest
One Facet G-Theory Model 1/1/2015 to 3/25/2020

Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**Expected Mean Square**	**Component Variance**	**Component % of Variance**
Persons (p)	33,314	$8.0e^6$	239	$\sigma^2_{pt,e} + n_t\sigma^2_p$	105.2	78.7
Time (t)	1	4840	4840	$\sigma^2_{pt,e} + n_p\sigma^2_t$	0.144	0.001
Residual (pt,e)	33,314	$9.4e^5$	28.3	$\sigma^2_{pt,e}$	28.3	21.1

The similarities continue in the Compliance data shown in Table 9.8. We again see a small positive variance for the time facet with more than 82% of the component variance accounted for in the Persons component. The percentage of variance attributed to the Time facet for the Compliance scale one of the lowest of the five scales at only 0.00%. This scale is once again in the same situation as both the Influence scale with more than 80% of the variance attributed to the Persons facet with approximately 26% unaccounted for.

9.2.2 One-Facet Model Reliability and Dependability

The end results of the G-theory one facet model applied to the TTI Success Insights Style Insights test-retest data is that the majority of the component variance is explained by the Persons

Table 9.8: Style Insights Compliance Scale Test-Retest
One-Facet G-Theory Model 1/1/2015 to 3/25/2020

Source of Variation	df	Sum of Squares	Mean Square	Expected Mean Square	Component Variance	Component % of Variance
Persons (p)	33,314	$8.8e^6$	264	$\sigma^2_{pt,e} + n_t\sigma^2_p$	119	82.1
Time (t)	1	400	400	$\sigma^2_{pt,e} + n_p\sigma^2_t$	0.386	0.00
Residual (pt,e)	33,314	$8.6e^5$	25.9	$\sigma^2_{pt,e}$	25.9	17.9

component rather than the Time component, although there is a non-trivial amount of component variance unaccounted for. It is worth noting once again that unaccounted for variance has traditionally been labelled as being due to random effects, yet this is not the interpretation when considering the implications of Generalizability theory. In fact, the unaccounted for variance is due to effects that have not been identified for one reason or another, not random effects that cannot be understood.

Finally, we present the generalizability and dependability coefficients developed in Section 9.1, see (9.7) and (9.9). Note that from the two previously mentioned equations, the relationship to computing the reliability and dependability coefficients is such that the main component of the computation is the value of σ^2_p compared to the δ and Δ errors presented in (9.6) and (9.8). The results are such that the generalizability and dependability coefficients are very close to being the same.

Table 9.9: Style Insights One-Facet Test-Retest Generalizability and Dependability
1/1/2015 to 3/25/2020

Scale	Generalizability	Dependability
D	0.89	0.89
I	0.91	0.91
S	0.88	0.88
C	0.90	0.90

It is noted that the Generalizability coefficient may be thought of as relatively synonymous to the standard reliability coefficient. It is not completely clear that there is a 1-1 relationship between the scoring levels as presented in Table 7.2. However, it is safe to interpret the coefficients in Table 9.9 in a similar fashion. With all scores at or above 0.84, all scales on the TTI Success Insights Style Insights assessment score quite well under the G-theory one facet model approach with Time as the facet.

As a final note, with all generalizability and dependability coefficients at or above 0.84 for all scales, the post-2015 version of the TTI Success Insights Style Insights assessment scales perform quite well under the one facet G-theory modeling approach with Time as the facet under consideration.

9.3 G-theory Summary

This section presented a brief introduction to the theoretical foundation of the one and two facet Generalizability theory models. We implement a one-facet model with time as the facet. The

results show that for all scales, time is not a factor in differences in test scores. The majority of the residual variances align with Persons, which is as desired.

10 Exploratory 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., [77], [84],[99], [100], 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.

10.1 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 [84], 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 [84], 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 [84] pp. 5.

According to [84] pp. 4,

> Thus, the chief aim is to attain scientific parsimony or economy of description.

On this topic of parsimony, [99] 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 [99] 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 [84] on pp. 97-98, where the author provides the following quotes from Thurstone's original work on the topic, see [169] 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)

To provide more context for the above quote, the following is also presented on pp. 98 of [84].

> 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 [84], although other authors are consulted.

10.2 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

3.3. 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 of real analysis from mathematics. 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

$$||a|| = \sqrt{\sum_{i=1}^{n} a_i^2}. \tag{10.1}$$

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

$$a \cdot b = \sum_{i=1}^{n} a_i b_i. \tag{10.2}$$

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.

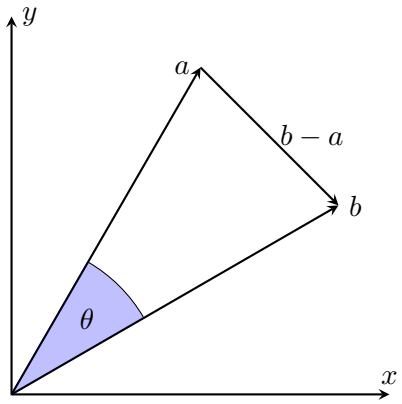

Figure 10.1: Vectors in $\mathbb{R}^n$

Proposition 10.1. *For any two vectors $a, b \in \mathbb{R}^n$, the angle θ between them is defined by the relationship*

$$\cos(\theta) = \frac{a \cdot b}{||a||||b||}. \tag{10.3}$$

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,

$$||b-a||^2 = ||a||^2 + |b||^2 - 2||a||||b|| \cos(\theta). \tag{10.4}$$

Using both (10.1) and (10.2), applied to the left hand side of the equality in (10.4), we have

$$||b-a||^2 = (b-a)\cdot(b-a) = b\cdot b - 2b\cdot a + a\cdot a = ||b||^2 + ||a||^2 - 2a\cdot b, \tag{10.5}$$

where we have used the commutative property of the scalar product in $\mathbb{R}^n$. Substituting the result of the final equality of (10.5) into (10.4), we have

$$||b||^2 + ||a||^2 - 2a\cdot b = ||a||^2 + |b||^2 - 2||a||||b|| \cos(\theta) \tag{10.6}$$

which implies

$$\cos(\theta) = \frac{a\cdot b}{||a||||b||}, \tag{10.7}$$

the result claimed in (10.3) 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

$$\mu(X) = E[X] = \sum_{i=1}^{n} X_i p_i. \tag{10.8}$$

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

$$\mu(X) = E[X] = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{10.9}$$

which is often denoted

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{10.10}$$

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

$$\sigma^2(X) = E[(X-\mu)^2] = \sum_{i=1}^{n} (X_i - \mu)^2 p_i. \tag{10.11}$$

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 and it 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 $\sigma^2(X)$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2. \tag{10.12}$$

Another possibility is to define $p_i = \frac{1}{n-1}$ and we have the so-called unbiased estimator

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2. \tag{10.13}$$

Adopting the usual convention that "pretty much everything may be considered normally distributed", we may consider the standardized version of a random variable X defined as

$$x = \frac{X - \mu}{\sigma}, \tag{10.14}$$

and we may note that $\mu(x) = 0$ and $\sigma(x) = 1$. We state this in the following proposition.

Proposition 10.2. *Let X be a random variable with mean μ and variance σ^2. Let $Z = \frac{X-\mu}{\sigma}$. Then, we have*

$$\begin{aligned} \mu(Z) &= 0 \\ &\textit{and} \\ \sigma(Z) &= 1. \end{aligned} \tag{10.15}$$

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,

$$\begin{aligned} E[Z] &= E[\tfrac{X-\mu}{\sigma}] \\ &= E[\beta] = \tfrac{1}{\sigma}E[X] - \tfrac{\mu}{\sigma} \\ &= \tfrac{\mu}{\sigma} - \tfrac{\mu}{\sigma} \\ &= 0. \end{aligned} \tag{10.16}$$

For the standard deviation of Z, it is convenient to first show that $\sigma^2(X)$ from (10.11) may be written in an alternative form.

$$\begin{aligned} E[(X-\mu)^2] &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X]^2 - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned} \tag{10.17}$$

Now, using the final line of (10.17) and the results of (10.16) we have

$$\begin{aligned} \sigma^2(Z) &= E[Z^2] - E[Z]^2 \\ &= E[(\frac{X-\mu}{\sigma})^2] - 0^2 \end{aligned}$$

$$\begin{aligned} &= \tfrac{1}{\sigma^2}E[X^2 - 2X\mu + E[X]^2] \\ &= \tfrac{1}{\sigma^2}(E[X^2] - 2E[X]\mu + E[X]^2) \\ &= \tfrac{1}{\sigma^2}(\{E[X^2] - E[X]^2\} + \{E[X]^2 - 2E[X]\mu + E[X]^2\}) \\ &= \tfrac{1}{\sigma^2}(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2) \\ &= \tfrac{1}{\sigma^2} * \sigma^2 \\ &= 1. \end{aligned} \tag{10.18}$$

□

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

$$\sigma(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \frac{1}{n}\sum_{i=1}^{n}(X - \mu_X)(Y - \mu_Y). \tag{10.19}$$

We have used the notation μ_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 (10.19) is to normalize it via the following approach

$$\rho_{X,Y} = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}}. \tag{10.20}$$

If we standardize both X and Y in (10.20) using (10.14), we may write

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{10.21}$$

Now, from (10.1) and (10.2), we may write (10.21) as

$$\rho_{X,Y} = \frac{x \cdot y}{||x||||y||}. \tag{10.22}$$

Finally, from (10.3) we have

$$\rho_{X,Y} = \frac{x \cdot y}{||x||||y||} = \cos(\theta), \tag{10.23}$$

where θ is the angle between x and y in $\mathbb{R}^n$.

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

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}}. \tag{10.24}$$

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

$$p = 2\int_0^t f(x)dx, \tag{10.25}$$

where t is the value computed in (10.24) and $f(x)$ is the probability density function of the t-distribution. For completeness,

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{10.26}$$

where ν represents the degrees of freedom and Γ is the gamma-function

$$\Gamma(n) = (n-1)! \tag{10.27}$$

for any positive integer n and

$$\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)dx \tag{10.28}$$

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 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

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{10.29}$$

and we also consider the (signed) square root of (10.29)

$$R = \pm\sqrt{R^2} = \sigma(\hat{Y}, Y). \tag{10.30}$$

10.3 The Basic Factor Model

We begin with a description of the basic factor model and some of its assumptions. In the social sciences, we assume 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 [100] 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 [100] posits the existence of two factors to explain this data, without naming these potential factors.

Table 10.1: Examples of Observed Variables from [100]

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 10.2 and create standardized variables Z_j rather than considering the set of X_j. The reason for this is pointed out in [84] 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 10.2. Once the variables are standardized, we may then form what [84] calls the *Full Classical Factor Model*,

$$z_{ji} = \sum_{k=1}^{m} a_{jk} F_{ki} + u_j Y_{ji}. \tag{10.31}$$

In (10.31) 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

$$cov\left(F_{ki}, Y_{ji}\right) = 0, \quad \forall i. \tag{10.32}$$

Further, it is assumed that the unique factors are uncorrelated with each other,

$$cov\left(Y_{jk}, Y_{ji}\right) = 0, \quad \forall j. \tag{10.33}$$

If we apply (10.12) to 10.31, noting again that z_{ij} are standardized variables with zero mean and unit variance, we obtain

$$\begin{aligned} s_j^2 &= \sum_{i=1}^{N} z_{ij}^2/N \\ &= \sum_{k=1}^{m} a_{jk}^2 \left(\sum_{i=1}^{N} F_{ki}/N\right) + u_j^2 \sum_{i=1}^{N} \left(Y_{ji}^2/N\right) \\ &+ 2 \sum_{p<q=1}^{m} a_{jp} a_{jq} \left(\sum_{i=1}^{N} F_{pi} F_{qi}/N\right) + 2u_j \sum_{j=1}^{m} a_{jp} \left(\sum_{i=1}^{N} F_{pi} Y_{ji}/N\right). \end{aligned} \tag{10.34}$$

We interpret the terms in (10.34) as follows. The first line of the equation is the definition which is expanded in terms of the right hand side of (10.31). 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 (10.34) 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 (10.34) as

$$s_j^2 = \sum_{i=1}^{m} a_{ji}^2 + u_j^2. \tag{10.35}$$

In (10.35), 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

$$V_i = \sum_{j=1}^{n} a_{ji}^2, \tag{10.36}$$

and further, define

$$V = \sum_{i=1}^{m} V_i \tag{10.37}$$

to be the total contribution of all common factors to the total variance. Then, V/n is the ratio of 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

$$h_j^2 = \sum_{i=1}^{m} a_{ji}^2 \quad j = 1, \ldots, n \tag{10.38}$$

is called the communality and the quantity

$$u_j^2 = 1 - h_j^2 \tag{10.39}$$

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.

$$z_j = \sum_{i=1}^{m} a_{ji} F_i + b_j S_j + e_j E_j. \tag{10.40}$$

In (10.40), 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 [84]. 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 [84]. 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.

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 [73]. See also Section 7 earlier 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 [84].

10.4 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, (10.31) becomes

$$Z = Af + Uy, \tag{10.41}$$

where we choose to follow the notation of [84] 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 (10.41) 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

$$Z = AF. \tag{10.42}$$

If we post-multiply (10.42) by F' (since F is not square, we must use the transpose) to obtain

$$ZF' = AFF' = S. \tag{10.43}$$

Note that the far left hand side of (10.43) 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

$$\tilde{R} = AFF'A' = A\Phi A' = S\Phi^{-1}S'. \tag{10.44}$$

10.5 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 [84], 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 [169]. This approach is the alternative offered to, and not to be confused with, maximizing the variance extracted as in the principal factors approach. While we do not directly discuss this concept in this section, one may consult Section 13.5 where we give a brief introduction to principal component analysis. One may also consult any of a number of texts and articles on the topic, such as Chapter 8 in [84].

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

$$\min_{A} ||R - I - (AA' - diag(AA'))||. \tag{10.45}$$

Now, explicitly writing the objective function for the $\frac{n(n-1)}{2}$ off-diagonal residual correlations

$$f(A) = \sum_{k=j+1}^{n} \sum_{j=1}^{n-1} \left(r_{jk} - \sum_{i=1}^{m} a_{ji} a_{ki} \right)^2. \tag{10.46}$$

It is noted in [84] 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

$$g(A) = \sqrt{\frac{2f(a)}{n(n-1)}}. \tag{10.47}$$

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 (10.45) 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 (10.46) or (10.47). 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

$$h_j^2 = \sum_{i=1}^{m} a_{ji}^2 \leq 1. \tag{10.48}$$

The authors do not intend to give a complete proof of the minimizer of (10.46) subject to the constraints (10.48), 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 an 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 [84] that the basic factor theorem given in equation (2.50) on pp. 32 of [84] 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

$$\hat{r}_{jk} = \sum_{i=1}^{m} a_{ki} b_{ji}. \tag{10.49}$$

We now compute the j^{th} sum of squares of off-diagonal residuals as

$$\begin{aligned} f_j(A) &= \sum_{\substack{k=1 \\ k\neq j}}^{n} \left(r_{jk} - \sum_{i=1}^{m} a_{ki} b_{ji} \right)^2 \\ &= \sum_{\substack{k=1 \\ k\neq j}}^{n} \left(r_{jk}^* - \sum_{i=1}^{m} a_{ki} \epsilon_i \right)^2, \end{aligned} \tag{10.50}$$

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 (10.50) with respect to the displacement ϵ

$$\frac{\partial f_j}{\partial \epsilon_l} = 2 \sum_{\substack{k=1 \\ k\neq j}}^{n} \left(r_{jk} - \sum_{i=1}^{m} a_{ki} e_i \right) (-a_{kl}). \tag{10.51}$$

Finally, we set the right hand side of (10.51) equal to zero and solve for the terms containing ϵ_i. After some rearranging, which is justified as all sums are finite, we obtain

$$\sum_{\substack{k=1\\k\neq j}}^{n} r_{jk}a_{kl} = \sum_{i=1}^{m}\left(\sum_{\substack{k=1\\k\neq j}}^{n} a_{ki}a_{kl}\right)\epsilon_i. \tag{10.52}$$

The final solution is most readily demonstrated in matrix form as

$$\epsilon_j \bar{A}_j' \bar{A}_j = r_j^0 A \tag{10.53}$$

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

$$\epsilon_j = r_j^0 A \left(\bar{A}_j'\bar{A}_j\right)^{-1}. \tag{10.54}$$

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 [84]. 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 to address the last issue, the so-called Heywood case and the reason for the constraints (10.48).

Theorem 10.3. *If the minimum of f_j is attained at a point outside the constraint set defined by (10.48), then a minimum of f_j subject to (10.48) will be attained on the boundary so that (10.48) may be replaced with*

$$\sum_{i=1}^{m} b_{ji}^2 = 1 \tag{10.55}$$

where we have replaced the a_{ji}^2 factor loadings with b_{ji}^2 as in (10.49) (since they are the minimization variables).

Theorem 10.3 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 (10.48) defines a compact set. The function (10.46) 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 [84].

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

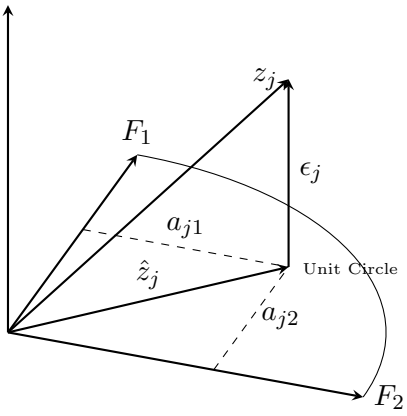

Figure 10.2: Geometric Interpretation of Factor Analysis Parameters

which minimizes the objective function given in (10.46) subject to the constraints (10.48) 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.

10.6 Factor Scores and Indeterminacy

In this section we discuss the problem of determinacy for the factor problem. Recall that in Section 10.4 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. Fro convenience, we now call this matrix $P = (A|U)$. Note that we many now write the system (10.41) in the form

$$Z = P(F|Y) = P\tilde{F} \tag{10.56}$$

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 10.5 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. [80] provides the example of a criterion that must be satisfied for an orthogonal solution

$$\frac{1}{n}FF' = I_n, \tag{10.57}$$

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. [80] 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 [80], 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 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

$$F = WZ \tag{10.58}$$

or, the score on factor j for person i may be written

$$F_{ji} = \sum_{k=1}^{N} w_{jk} z_{ik}. \tag{10.59}$$

The most simple form of a solution is to take $W = A'$ where A is the factor loading matrix from the P in (10.56). 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 (10.58). 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 [80]. 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.

11 A Graded Response Theory Approach to Likert Response Format DISC

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 the discussion in Chapter 12.

As noted earlier in this manual, the TTI Success Insights Style Insights assessment is a four-item, forced-rank style assessment. Such a response format may not be modeled, strictly speaking, by a polytomous response model. However, during the process of testing and updating the Style Insights assessment model, a host of data is to be collected and analyzed in a Likert-style format, using a six-point response format. It is the intent of this process to use the tools available to the suite of polytomous response models to analyze the Style Insights response data as part of the decision making process behind building the next generation of forced-rank frames for 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 [8], [130], and [133].

The article in [133] is a synopsis of polytomous IRT models and is part of a larger series on quantitative applications in the social sciences. The focus of [130] 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 [8] 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 today. 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.

11.1 From Classical Test Theory to Item Response Theory

Some authors attribute the foundation of Classical Test Theory to Cattel around 1890, see [36] and [133], while others trace it to the work of Spearman in the early part of the 20^{th} century, see [157]. 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

$$X = T + E, \tag{11.1}$$

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

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \tag{11.2}$$

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 analysis. 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 11.1.

Table 11.1: 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 11.1 and creates Table 11.2. Simply summing the values in Table 11.2 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.

Table 11.2: 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 [133], 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 [83] and [133], one gets a feel that there are slightly differing opinions of the evolution of IRT over the years. For example, [133] traces the foundation of IRT to Fechner circa 1860 (see [7]) and Thurstone in 1925 (see [14], [168], and [175]). The mathematical roots of IRT in psychology may be traced back to the work of Binet, Simon, and Terman from 1916 (see [7] and [14]), while the formal basis of IRT is generally attributed to Lawley for his work in 1943 (see [7], [104], and [175]). Lord (see [115]) formalized IRT as the extension of CTT (see [7]). Lord ([116]), Lord & Novick ([117]), and Birnbaum ([13]) helped establish the acceptance of IRT in the psychological community. Finally, Rasch (see [139], [140], and [141]) provided the first specific class of IRT models.

Compare the presentation of [83], which discusses more of the contemporary development of IRT beginning with Lord (see [113], [114], and [115]), citing that Lord led the transition from CTT to IRT from Gulliksen's (see [81]) challenge to develop invariant item statistics for test development. [83] also notes the development of the "item characteristic curve" (see [170]) and the use of the term "latent trait" (see [105]). Finally, a suggestion to change the terms "item characteristic curve" and "trait" to "item response function" and "ability" is provided by Lord (see [116]).

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, [133] provides a somewhat more motivation based approach, while on the other, [83] 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.

11.2 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 (negatibve) 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 11.1.

There are some interesting characteristics of a CDF that have natural interpretations in item analysis. Note first that the CDF displayed in Figure 11.1 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 $\mathbb{R}$.

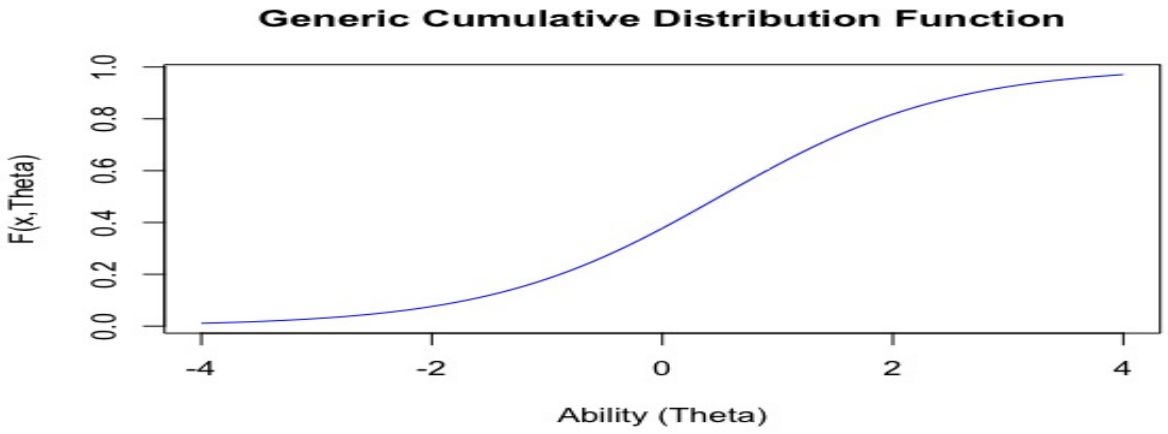

Figure 11.1: Generic CDF

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.

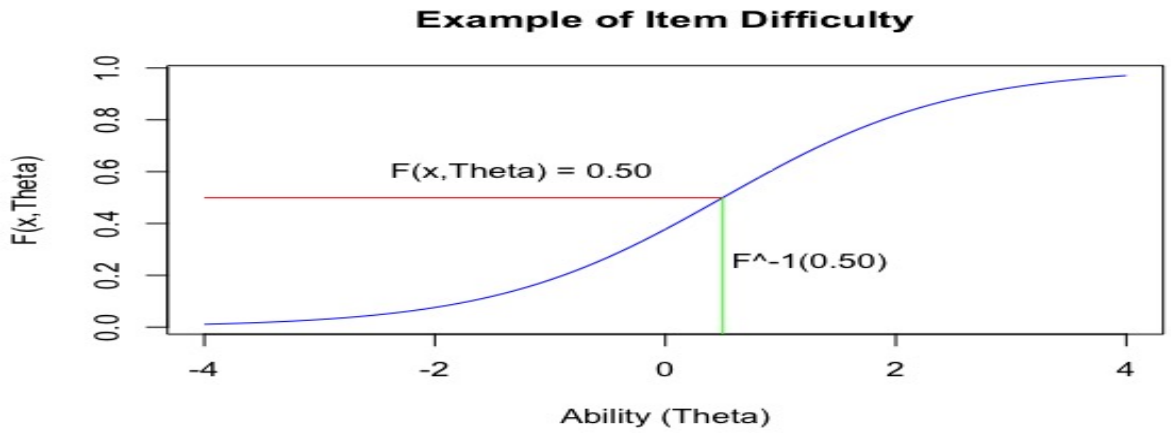

Figure 11.2: Example of Item Difficulty

Figure 11.2 is a continuation of Figure 11.1 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

equation $F(x) = y$, where y is the midpoint between the lower and upper asymptotes of the CDF. In most, if not all, cases this solution is likely to exist by invoking strict monotonicity of the CDF.

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 to 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 CDF attains a global maximum, x_0. See Figure 11.3 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.

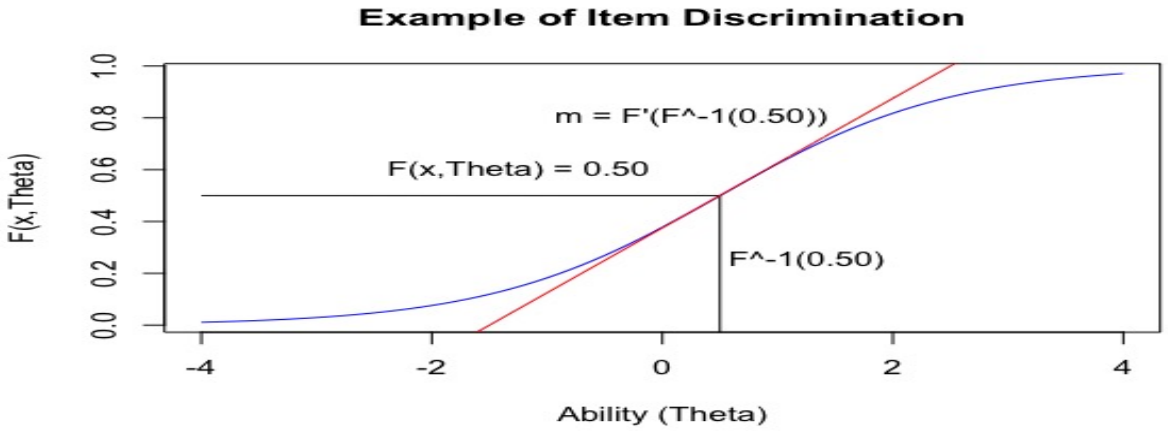

Figure 11.3: Example of Item Discrimination

Definition 11.1. *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 the multitude of available advanced calculus or introductory real analysis text books for a discussion of these and other similar topics.

Theorem 11.2. *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 11.3. *Every continuous 1-1 real-valued function on an interval is strictly monotone.*

Theorem 11.4. *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 [139]. There are many equivalent ways to write down this model, so we choose the following notation:

$$Pr\{X_{ni} = 1\} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \tag{11.3}$$

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.

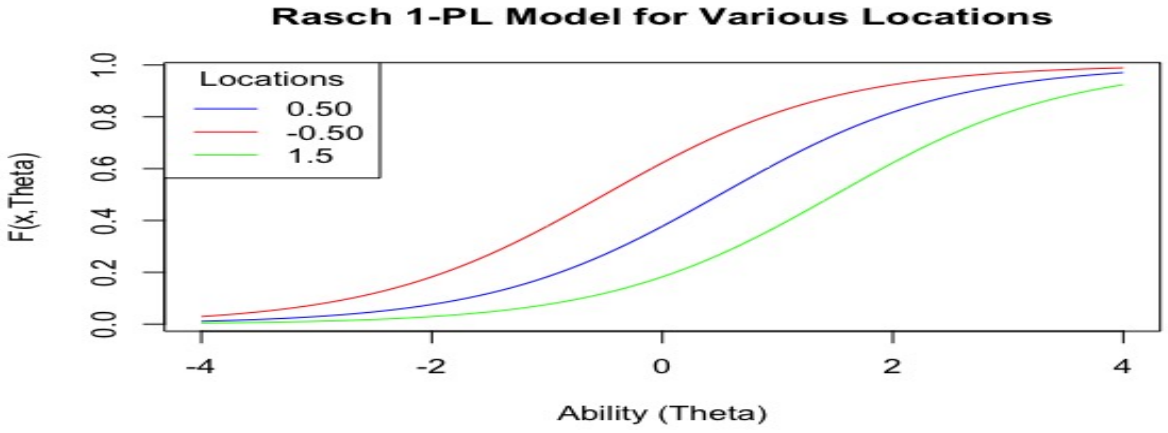

Figure 11.4: Rasch 1-PL Model

Figure 11.4 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 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.

$$f(x) = \frac{\exp\left(a\left(x-\delta\right)\right)}{1+\exp\left(a\left(x-\delta\right)\right)}, \tag{11.4}$$

where a now represents the discrimination parameter.

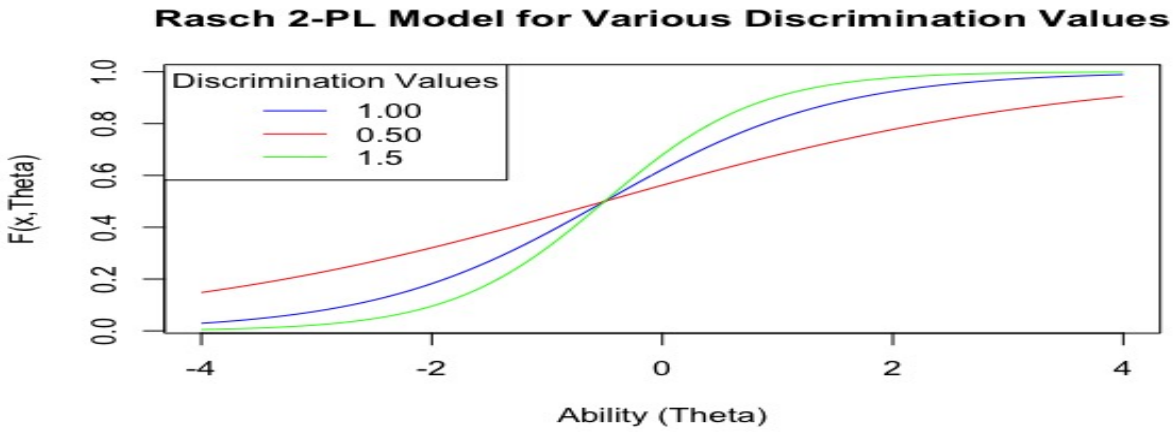

Figure 11.5: Rasch 2-PL Model

Figure 11.5 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 (11.4) as follows:

$$f(x) = c + \frac{(1-c)\exp\left(a\left(x-\delta\right)\right)}{1+\exp\left(a\left(x-\delta\right)\right)}. \tag{11.5}$$

It is instructive to spend a moment exploring the formula in (11.5). Note that if we set $c = 0$ we recover exactly the formula in (11.4) for the 2-PL model. Similarly, if we set $c = 0$ and $a = 1$ in (11.5), we recover the formula in (11.3) for the 1-PL model. Figure 11.6 shows an example of the 3-PL model with $c = 0.25$.

There is a 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 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.

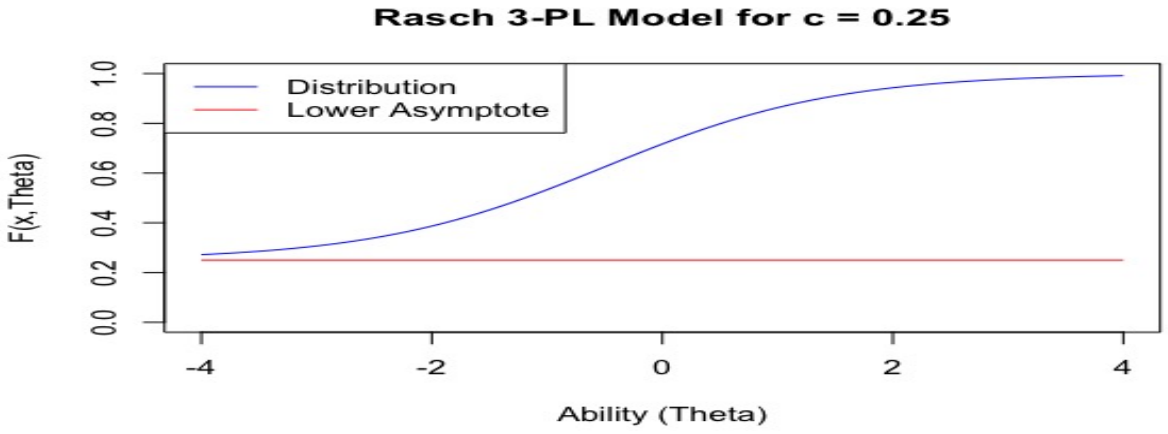

Figure 11.6: Rasch 3-PL Model

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.

11.3 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 [133], 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. 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 [133], pp. 63:

> In Samejima's framework, however, the plausibility of the response process thought to generate the data is paramount, because for Samejima ([149]), 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 of interest. 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 [148], [149]) demonstrates the increase in information available when gathering the information with more categories (polytomous), while Cohen (see [45]) 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 [133], 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 [133]. 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 [133], 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

$$P_{i_j} = P^*_{i_j} - P^*_{i_{j+1}}, \tag{11.6}$$

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 [130] 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 (11.6). 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 [133]. 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, [146];
2. GRM is the polytomous manifestation of Thurstone's method of successive intervals, see [35], [62], and [144];
3. GRM relies on (Rasch) 2-PL model, see e.g., [54], [68], [69], and [151], to obtain cumulative boundary functions $P^*_{i_j}$;
4. The functions $P^*_{i_j}$ could be built based on any appropriate mathematical function, see [147], [150];
5. In practice, only 2PL and 2-parameter normal ogive have been used; both outlined in Samejima's original work, see [146].

Following the lead of the 2-PL model given in (11.4) we define the category boundary response function (CBRF) of the graded response model (GRM) to be

$$P^*_{i_j}(x) = \frac{\exp\left(a_i\left(x - \beta_{i_j}\right)\right)}{1 + \exp\left(a_i\left(x - \beta_{i_j}\right)\right)}, \tag{11.7}$$

where the a_i represent the discrimination parameter of the i^{th} item and β_{i_j} 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 (11.6) 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 $P^*_{i_j}$ as provided for in (11.7). $P_{i_1} = P^*_{i_0} - P^*_{i_1} = 1 - P^*_{i_1}$, $P_{i_n} = P^*_{i_n} - P^*_{i_{n+1}} = P^*_{i_n}$, and the remainder are as defined in (11.6). To summarize,

$$P_{i_j} = \begin{cases} 1 - P^*_{i_1} & \text{if } \ j = 0, \\ P^*_{i_n} & \text{if } \ j = n, \\ P^*_{i_j} - P^*_{i_{j+1}} & \text{otherwise.} \end{cases} \tag{11.8}$$

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.

11.4 Item and Test Information

One key area that assessments are required to address is, as noted in Section 11.1, the concept of reliability. Typically, one uses a single metric such as the α coefficient of Cronbach, see [56]. More advanced treatments relying on a factor analytic approach may consider McDonald's ω, see [178]. 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 and test information functions 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 11.1. Recall the amended notation from that section and let $P_i(\theta)$ denote the probability model for the 1-PL model as shown in (11.3). Let $Q_i(\theta) = 1 - P_i(\theta)$. According to Fisher information theory [67], the information function for the 1-PL model is given by

$$\mathcal{I}_i(\theta) = P_i(\theta)Q_i(\theta). \tag{11.9}$$

Note that the item information function in this case is related to the item characteristic functions defined by $P_i(\theta)$. [8] gives an excellent introduction from the viewpoint of item characteristic functions. In particular, one can show that for the 1-PL model

$$\mathcal{I}_i(\theta) = \frac{\partial P_i(\theta)}{\partial \theta}. \tag{11.10}$$

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

$$\mathcal{I}(\theta) = \sum_{i=1}^{n} \mathcal{I}_i(\theta). \tag{11.11}$$

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

$$SE(\theta) = \frac{1}{\sqrt{\mathcal{I}(\theta)}}. \tag{11.12}$$

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 [8] for a more complete treatment.

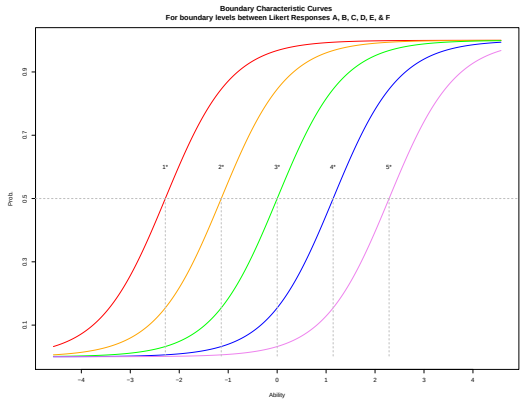

Figure 11.7: Example Boundary Characteristic Curves

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 11.3. Also noted in that section is that there exists 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 for one item for this type assessment.

Continuing to follow the development of the GRM model from Section 11.3, we define the item category response characteristic curves as previously outlined. The following is a theoretical example based on the boundary curves presented in Figure 11.7.

It is noted that the information presented in Figures 11.7 and 11.8 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 11.8. It is unlikely to expect such high quality results in practice. We present them as examples to be used for comparison purposes.

We now turn our attention to constructing the item and test information functions for the polytomous response case, specifically for the GRM. Following [8], we note that the ultimate goal is to determine the precision with which a test or assessment measures ability and this has traditionally

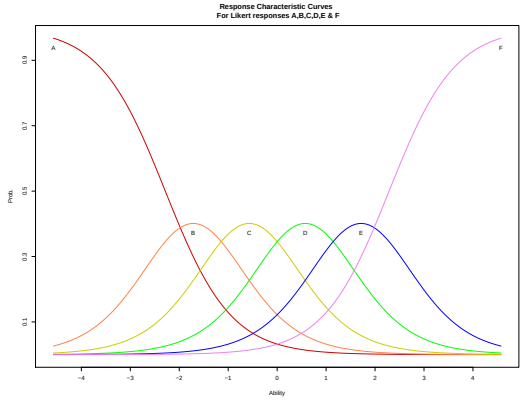

Figure 11.8: Example Item Response Category Characteristic Curves

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

$$I_i(\theta) = \sum_{j=1}^{n} I_{ij}(\theta)P_{ij}(\theta). \tag{11.13}$$

Relying on Samejima's development, see [146] and [147], define the information function of an item response category to be

$$I_{ij}(\theta) = -\frac{\partial^2 log(P_{ij}(\theta))}{\partial\theta^2} = -\frac{\partial}{\partial\theta}\left(\frac{P'_{ij}(\theta)}{P_{ij}(\theta)}\right), \tag{11.14}$$

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

$$I_{ij}(\theta) = \frac{\left(P'_{ij}(\theta)\right)^2 - P_{ij}(\theta)P''(\theta)}{(P_{ij})^2}. \tag{11.15}$$

We may then compute the share of the item information contributed by category j for item i to be

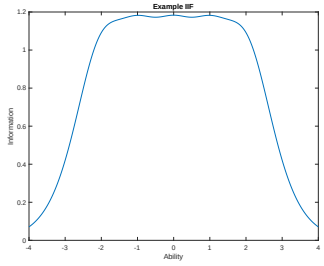

Figure 11.9: Example Item Information Function

$$I_{ij}(\theta)P_{ij}(\theta) = \frac{\left(P'_{ij}(\theta)\right)^2}{(P_{ij})^2} - P''_{ij}(\theta). \tag{11.16}$$

Finally, the item information for item i becomes

$$I_i(\theta) = \sum_{j=1}^{n} \left(\frac{\left(P'_{ij}(\theta)\right)^2}{(P_{ij}(\theta))^2} - P''_{ij}(\theta) \right). \tag{11.17}$$

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 [146]. In that case, the computation in terms of boundary probabilities reduces to

$$I_i(\theta) = \sum_{j=1}^{n} \frac{\left(P^{*\prime}_{ij-1}(\theta) - P^{*\prime}_{ij}(\theta)\right)^2}{P^*_{ij-1}(\theta) - P^*_{ij}(\theta)}. \tag{11.18}$$

Note that an interesting exercise is to show that (11.13) reduces to

$$I_i(\theta) = \frac{(P'_i(\theta))^2}{P_i(\theta)Q_i(\theta)} \tag{11.19}$$

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 (11.17), we have

$$T = \sum_{i=1}^{N} I_i = \sum_{i=1}^{N} \sum_{j=1}^{n} \left(\frac{\left(P'_{ij}\right)^2}{\left(P_{ij}\right)^2} - P''_{ij} \right), \tag{11.20}$$

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.

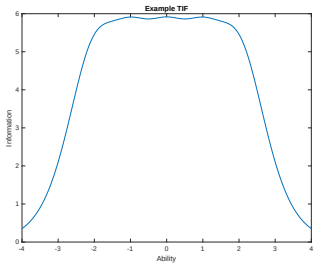

Figure 11.10: Example Test Information Function

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 (difficult). It may then seem reasonable to apply the typical lower boundaries for reliability coefficients as first presented in Table 7.2 in Section 7.2.1. 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 7.2.1.
An example is given in Figure 11.11 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.

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 11.12.

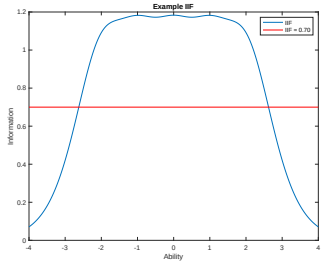

Figure 11.11: Example Item Information Function with Reliability Boundary

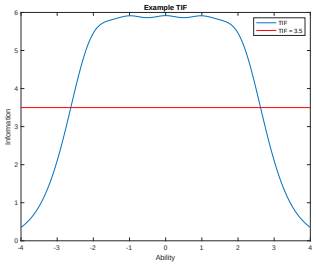

Figure 11.12: Example Test Information Function with Reliability Boundary

11.5 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 somewhat challenging to follow for the non-mathematically adept individuals. 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.

12 Thurstonian Item Response Theory Approach to Forced-rank Assessments

There appears to be a divide in the assessment world regarding the positive and negative consequences of choosing to use the likert-style or forced-choice style in self report response formats. Many of the tools developed over the decades are geared directly toward likert-style response formats. One of the main criticisms levied by the academic community toward assessment companies using the forced-choice response format is that there are no adequate tools to readily measure the evidence of validity and reliability of these style assessments. In defense of this criticism, many developers of the forced-choice response format assessments have routinely relied on the use of tools developed for likert-style response format assessments claiming valid and reliable assessments based on the positive outcomes of such analyses.

As an example, many developers and users of forced-choice response format assessments use and rely upon the results of the α coefficient of Cronbach, see [56], to measure the reliability of their assessment scales. It is rarely mentioned that the α coefficient is not a reliable tool to measure the reliability of such scales as it does not, and cannot, account for the effects of the assessment takers responses to the scale in question and its interaction with the remaining scales. This does not imply that an assessment scale lacks internal consistency. Rather, such analyses cannot establish that the scale has evidence of internal consistency, at least not to the academic and profession standards in the assessment community today.

Some of the arguments against the forced-choice format are as follows. The assessment taker is forced to respond to the item, whether they have an opinion or not. The forced-choice format can only be used to measure an individual and should not be used to compare individuals. It is not possible to capture the level of intensity an individual has on a scale using forced-choice formats. Poorly developed forced-choice assessments inherently introduce a bias into the assessment output. These are just a few examples of the criticisms of this style assessment response format.

The current authors would argue that each of the previous arguments has merit. We would also argue that several of the same arguments apply nearly as equally to a likert-style response format. For example, there is a claim that likert-style response formats allow for the assessment to be used to measure intensity. However, how well can it measure intensity without relying on certain assumptions? The main assumption that must be made is that the scale items all equally measure the scale and that distance units of one are adequate to the task of capturing the aforementioned intensity. It is hard to imagine these assumptions hold in practice.

Classical test theory (CTT) is based on assumptions that likely do not hold in practice and many assessment companies rely solely on CTT to measure the evidence of validity and reliability of their assessment scales, at least according to many of the technical manuals available from these companies today. Since the purpose of this section is to establish evidence of the validity of the TTI Success Insights Style Insights assessment, and not to attack our competitors, we leave it to the reader to discover the truth of the previous statement. Many companies have gone away from the forced-choice format with their versions of the DISC model. Many Big Five assessments are in the likert-style format. Given that there remain questions surrounding this format, one would expect the developers and users of such assessments in the likert-style format to search for more adequate analyses.

One does not need to venture far into the assessment literature to find item response theory (IRT). In particular graded response models (GRM) are a quite useful tool for analyzing likert-style assess-

ment format data. One may consult Section 11 of this manual for a discussion of the TTI Success Insights approach to GRM as applied to likert-style versions of our DISC model for analytic purposes. In that section, we use the difficulty and discrimination parameters of the GRM output to establish grouping of items from the different scales that show similar patterns. This is an attempt to minimize the bias that may be introduced in the assessment development process noted in an earlier part of this section.

Another of the arguments against the forced-choice format is that the assessment taker is forced to respond to the assessment item, whether they have an opinion or not. The likert-style format can allow for a neutral option. If one is measuring opinion, say of service received at a fine restaurant, then a neutral option may be justified. However, if one is attempting to measure an ability spectrum using a psychometric assessment, perhaps the researcher requires the respondent to choose.

Additionally, the likert-style format allows the individual to be all things at all times, if they choose. This is particularly true in the case where an assessment may be used in high stakes situations, such as selection for hiring or promotion. If the assessment taker is aware of the stakes, they may attempt to be exactly that, all things at all times. A forced-choice format does not allow for this type response pattern. Perhaps an example is in order.

Suppose we are interested in measuring what is really important to an individual that helps guide them in their lives. In other words, what is the main motivation behind the individual. Suppose we ask the open-ended question

The best return on any investment is ...

and we provide the following possible answers:

- the knowledge I gain from the investment.
- growing closer in my understanding of the world.
- recognition.
- the difference I make in others' lives.
- understanding how it helps me be more connected to my surroundings.
- money.

If we were to ask this in a forced-choice format, the individual is then asked to choose those outcomes that are most important to them, then next most, and so on. In a likert-style format, the individual may be able to rank several of these responses as all or nothing. Is it possible for an individual to be equally motivated by intellectual pursuit and power or money? We provide an anecdotal response to that, which is clearly a rare occurrence, but an occurrence nonetheless.

In August 2006, Grigori Perelman was awarded the Fields Medal for his contributions to the proof in the affirmative of the Poincaré conjecture, a century old problem first proposed by Henri Poincaré in 1904. Pereleman famously refused the medal as noted in a NY Times article from August 22, 2019. According to the NY Times article, Perelman met with Sir John M. Bald, president of the International Mathematical Union where he rejected again the award. According to New Yorker magazine in an article dated March 19, 2011, Perelman is quoted as saying

> He proposed to me three alternatives: accept and come; accept and don't come, we will send you the medal later; I don't accept the prize. From the very beginning, I told him I have chosen the third one ... [the prize] was completely irrelevant for me. Everybody understood that if the proof is correct, then no other recognition is needed.

In 2010, Perelman was further awarded the Millennium Prize, a one million dollar award given for the solution to one of the eight, at the time, major problems that existed when the Millennium Prize Problems were first established by the Clay Institute in 2000. The original problems are the Birch and Swinnerton-Dyer conjecture, the Hodge conjecture, existence and smoothness of Navier-Stokes equations, P versus NP problem, the Poincaré conjecture, the Riemann hypothesis, and the Yang-Mills existence and mass gap. To date, only Perelman's proof of the Poincaré conjecture has been awarded a Millennium Prize. According to [145], Perelman refused the award due to the lack of fairness surrounding the award, noting that his own contribution was no greater than that of Richard Hamilton. Further,

> To put it short, the main reason is my disagreement with the organized mathematical community. I don't like their decisions, I consider them unjust.

This is an individual who, according to some accounts lived in his parents' basement, who turned down the highest award and largest monetary prize in the field of mathematics because "the proof is enough" and for fairness. This is also an individual that is likely as high on the intellectual scale as one can be. This example is intended to illustrate that the highest level of a scale is very difficult to achieve and likely does not play well with other scales on the same assessment. This is also meant to be an argument, albeit anecdotal in nature, that in some cases we need to force choices on individuals to obtain true measures of scales.

The entire point established just now is completely moot if we lack the ability to measure evidence of validity and reliability adequately for the forced-choice format assessment. That is the main purpose of this section of this manual. For discussion, we present again some information from Section 3.5.

One of the main reasons that CTT is not applicable to forced-choice format data is the lack of existence of an inverse for the scale correlation or covariance matrix. In other words, many of the techniques of CTT, such as (exploratory or confirmatory) factor analysis require the inversion of a certain matrix during the solution for the parameters of the underlying nonlinear optimization problem. Essentially, most regression analyses rely on the inversion of a matrix on some level. In the case of many CTT techniques, this matrix is covariance matrix of the underlying assessment response data. In the case of forced-choice assessment data, the covariance matrix is, by definition, singular. This means that the matrix of covariation for a forced-choice assessment cannot be inverted.

As presented in Section 3.5, the authors in [94] note the following related to the correlation between scales on a forced-choice format assessment.

> 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 [94] 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 [94] 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.

We note that this article was first published in the Journal of Occupational Psychology in 1988. There have been some interesting developments in the years that lead to the writing of this manual.

12.1 The Basics of Thurstonian Item Response Theory Models

There are many ways to gather information related to human behavior through self-reporting assessments. Mentioned in the previous section were only two of the possibilities. Other options include distributing a fixed number of points between the options, resulting in compositional data, and attaching a preference scale to items allowing the respondent to indicate how much more or less they prefer one item to another. This last option is the topic of a paper on factor analyzing so-called graded preference format data, see [33].

Several papers by the same authors as in [33] have proposed variants of the Thurstonian item response theory approach to forced-choice assessments to allow for an IRT based scoring rubric that facilitates inter-individual comparisons, one of the main areas of concern noted in [94] and summarized in the previous section. These include, but are not limited to, [29], [30], [31], [32], [122], and [123].

We begin with a synopsis of the modeling approach, generally following that presented in [121]. Considering the forced choice approach to collecting information on individual behavior, suppose we are considering n objects from which we must make comparisons among all possible pairs of those objects. The reason for this comes from the Thurstonian Law of Comparative Judgment, which essentially states that in order for an individual to rank a set of n objects, that individual must make all possible comparisons between those objects. This leads to the idea of a combinatorial approach to the number of comparisons. In the language of combinatorics, the total number of comparisons of n objects taken two at a time is given by

$$\frac{n!}{(n-2)!2!} = \frac{n(n-1)(n-2)!}{(n-2)!2!} = \frac{n(n-1)}{2}. \tag{12.1}$$

Following [121], we assume that (a) no equality judgments are allowed, (b) each subject in the multiple sample is asked to judge all pairs, and (c) presentation order effects are negligible. Continuing along the lines of Thurstone's arguments, a pair of stimuli presented to a subject elicits a continuous preference or utility function. Further, the stimulus whose utility is larger at the moment of comparison is necessarily preferred by the subject. Finally, it is assumed that these unobservable preferences are normally distributed (with mean 0 and standard deviation 1).

Noting that the notation to come may be considered a bit cumbersome to some, we present without proof the following concepts. The outcomes of each paired comparison are presented as a binary indicator or dichotomous variable which we denote y_{lj}. y_{lj} denotes the response of subject j to comparison $l = (i, i')$ which is given as follows:

$$y_{lj} = \begin{cases} 1 & \text{if subject } j \text{ chooses object } i \\ 0 & \text{if subject } j \text{ chooses object } i' \end{cases} \tag{12.2}$$

where $\tilde{n}$ denotes the total number of choices presented to each subject j, $l = 1, \ldots, \tilde{n}$, $j = 1, \ldots, N$, $i = 1, \ldots, n-1$, and $i' = 1, \ldots, n$.

Letting t_{ij} denote the j^{th} subjects continuous preference for object i, Thurstone proposed the following transformation, y_l^*, of the unobserved preferences

$$y_l^* = t_i - t_{i'}. \tag{12.3}$$

The previous definition shows that under the preceding assumptions, the model chooses object i if $t_{ij} \geq t_{i'j}$ and chooses i' if $t_{ij} < t_{i'j}$. Combining this idea along with (12.2) and (12.3) implies that

$$y_{lj} = \begin{cases} 1 & \text{if } t_{ij} \geq t_{i'j} \\ 0 & \text{if } t_{ij} < t_{i'j} \end{cases}. \tag{12.4}$$

We next define the following variables. Let $\boldsymbol{t}$ be the n dimensional vector containing the n unobservable preferences. Let $\boldsymbol{e}$ denote the $\tilde{n} = \frac{n(n-1)}{2}$ dimensional vector of the random errors associated with the $\tilde{n}$ preference comparisons. We follow [161] in assuming that $\boldsymbol{t}$ is multivariate normally distributed with mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$. Further, assume that $\boldsymbol{e}$ is multivariate normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$. Finally we assume the following:

$$\begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{e} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_t \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_t & \\ \mathbf{0} & \boldsymbol{\Omega} \end{pmatrix} \right) \tag{12.5}$$

where $\boldsymbol{\Omega}$ is assumed diagonal with entries ω_l and the $\mathbf{0}$ in the joint covariance matrix in (12.5) is assumed to be of the appropriate size and represents the assumption that the covariances between the unobservable preferences and the errors associated with their pairwise comparisons are all 0.
If we assume the pairwise comparison given in (12.3), we may write the following relationship between the vector $\boldsymbol{y}^*$ and the left hand side of (12.5). Let A denote the $\tilde{n} \times n$ matrix defined in the following way

$$a_{lk} = \begin{cases} 0 & \text{if } k \notin \{i, i'\} \\ 1 & \text{if } k = i \\ -1 & \text{if } k = i' \end{cases}. \tag{12.6}$$

where the element a_{lk} is in the l^{th} row and k^{th} column of A. We may now write the left hand side of (12.5) using the following notation

$$\boldsymbol{y}^* = (\boldsymbol{A}\ \boldsymbol{I}) \begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{e} \end{pmatrix} = \boldsymbol{A}\boldsymbol{t} + \boldsymbol{e}, \tag{12.7}$$

where $\boldsymbol{I}$ is the appropriately sized identity matrix. As an example, consider the case of $n = 4$, corresponding to a single frame on the TTI Success Insights Style Insights assessment, a forced-choice frame with 4 items. In this case, the matrix A takes on the form

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{12.8}$$

According to [121], combining the definition of A and the definitions of y_{lj} from (12.4), we may write for any paired comparison pattern

$$Pr\left[\bigcap_{l=1}^{\tilde{n}} y_l\right] = \int_R \cdots \int \phi_{\tilde{n}}(\boldsymbol{y}^* : A\boldsymbol{\mu}_t, A\boldsymbol{\Sigma}_t A' + \boldsymbol{\Omega}) d\boldsymbol{y}^* \tag{12.9}$$

where $\phi_{\tilde{n}}(\cdot)$ denotes the $\tilde{n}$ dimensional normal density function with limits of integration

$$R_l = \begin{cases} (0, \infty) & \text{if} \quad y_l = 1 \\ (-\infty, 0) & \text{if} \quad y_l = 0 \end{cases} . \tag{12.10}$$

Ignoring the error term in (12.7), we simply write

$$\boldsymbol{y}^* = \boldsymbol{A}\boldsymbol{t} \tag{12.11}$$

and we move into the realm of the Thurstonian factor model by considering the latent utilities to be a (linear) function of the underlying traits of interest (in our case, the D, I, S, and C traits). In other words, following [30] write

$$\boldsymbol{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}. \tag{12.12}$$

This should come as no surprise as we noted in (12.5) that the latent utilities had a mean vector $\boldsymbol{\mu}_t$. We are merely writing the latent utilities as a linear function as a "deviation" from their means in some sense. If there are n means and d latent traits, then $\boldsymbol{\Lambda}$ is an $n \times d$ matrix of factor loadings, $\boldsymbol{\eta}$ is a $d\times$ vector of common factors (latent traits), and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of unique factors. Note that we make some assumptions that are common in a standard factor analytic approach. The latent traits ($\boldsymbol{\eta}$) are uncorrelated with the unique factors ($\boldsymbol{\epsilon}$) and have means of zero. The latent traits are freely correlated (their covariance matrix is denoted $\boldsymbol{\Phi}$), but variances are fixed to one to be able to identify. The unique factors are uncorrelated implying their covariance matrix ($\boldsymbol{\Psi}^2$) is diagonal. Finally, we assume that both the latent traits and unique factors are normally distributed.

At this point, the authors in [30] suggest reparameterizing the above model using equations (12.11) and (12.12). They give several reasons for this which we summarize in the following list.

1. The Thurstonian factor model presented measures second order factors, such as latent traits, versus first order factors, such as latent utilities
2. The Thurstonian IRT model speeds up computations in the case of large models
3. The residual error variances of the latent responses ($\boldsymbol{y}^*$) are zero for the factor model, implying latent trait estimates cannot be computed
4. Reparameterization of the factor model as a first order model produces nonzero residual error variances, allowing the computation of latent trait estimates.
5. Reparameterization allows for some important descriptors of IRT models to be formulated (item characteristic functions, item information functions, etc.)

We now reparameterize the model by using equations (12.11) and (12.12) as

$$\boldsymbol{y}^\star = A\boldsymbol{t} = A(\boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}) = A\boldsymbol{\mu}_t + A\boldsymbol{\Lambda}\boldsymbol{\eta} + A\boldsymbol{\epsilon} \tag{12.13}$$

and define the following terms. Let $-A\boldsymbol{\mu}_t = \gamma$, $A\boldsymbol{\Lambda} = \breve{\boldsymbol{\Lambda}}$, $A\boldsymbol{\epsilon} = \breve{\boldsymbol{\epsilon}}$, and $cov(\breve{\boldsymbol{\epsilon}}) = \breve{\boldsymbol{\Psi}}^2$. Note that $\breve{\boldsymbol{\Psi}}^2$ may be written

$$\breve{\mathbf{\Psi}}^2 = A\mathbf{\Psi}^2 A'.$$

Note that if we set p to be the number of frames or blocks of the n items we have the following.

1. $\gamma = -A\boldsymbol{\mu}_t$ is a vector of $(p\tilde{n}) \times 1$ (unrestricted) thresholds
2. $\breve{\mathbf{\Psi}}^2$ is the $(p\tilde{n}) \times (p\tilde{n})$ covariance matrix of the unique pairwise errors
3. $\breve{\mathbf{\Lambda}}$ is the $(p\tilde{n}) \times d$ matrix of factor loadings

To put some perspective on the scope of the problem for the TTI Success Insights Style Insights assessment, we have $p = 24$ blocks measuring $d = 4$ traits with $n = 4$ items per block. The matrix shown in (12.14) covariance relationships of the unique pairwise errors for a single four-item, forced-choice frame or block.

$$\breve{\mathbf{\Psi}}^2 = \begin{pmatrix} \psi_1^2 + \psi_2^2 & & & & & \\ \psi_1^2 & \psi_1^2 + \psi_3^2 & & & & \\ \psi_1^2 & \psi_1^2 & \psi_1^2 + \psi_4^2 & & & \\ -\psi_2^2 & \psi_3^2 & 0 & \psi_2^2 + \psi_3^2 & & \\ -\psi_2^2 & 0 & \psi_4^2 & \psi_2^2 & \psi_2^2 + \psi_4^2 & \\ 0 & \psi_3^2 & \psi_4^2 & -\psi_3^2 & \psi_4^2 & \psi_3^2 + \psi_4^2 \end{pmatrix} \tag{12.14}$$

In order to illustrate some of the structure, we consider a condensed example of an instrument similar to the Style Insights assessment containing $p = 3$ blocks. In this case, we first illustrate, for convenience, the case with 3 latent traits measured by 3 items per block. In this case, the matrix A is a block diagonal matrix of size 9×9 where each block diagonal entry is given by

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}.$$

We present the full matrix for this example to present the structure.

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \tag{12.15}$$

Continuing, the Λ matrix is a matrix of 3×3 matrices "stacked" vertically with each 3×3 "block" of the form

$$\begin{pmatrix} \lambda_i & 0 & 0 \\ 0 & \lambda_{i+1} & 0 \\ 0 & 0 & \lambda_{i+2} \end{pmatrix}$$

where in this example $i \in \{3j + 1 | j = 0, 1, 2\}$ and in general $i \in \{3j + 1 | j = 0, 1, 2, \dots\}$. The final Λ matrix for this example looks like

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ \lambda_4 & 0 & 0 \\ 0 & \lambda_5 & 0 \\ 0 & 0 & \lambda_6 \\ \lambda_7 & 0 & 0 \\ 0 & \lambda_8 & 0 \\ 0 & 0 & \lambda_9 \end{pmatrix} \tag{12.16}$$

We now compute the matrix $\breve{\mathbf{\Lambda}} = A\mathbf{\Lambda}$.

$$\breve{\Lambda} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & \lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & \lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & \lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \end{pmatrix} \tag{12.17}$$

Note that $\mathbf{\Psi}^2$ is the diagonal matrix with ψ_i^2 as the i^{th} diagonal entry. We then compute $\breve{\mathbf{\Psi}}^2 = A\mathbf{\Psi}^2A'$ to obtain

$$\breve{\mathbf{\Psi}}^2 = \begin{pmatrix} \psi_1^2+\psi_2^2 & & & & & & & & \\ \psi_1^2 & \psi_1^2+\psi_1^2 & & & & & & & \\ -\psi_2^2 & \psi_3^2 & \psi_2^2+\psi_3^2 & & & & & & \\ 0 & 0 & 0 & \psi_4^2+\psi_5^2 & & & & & \\ 0 & 0 & 0 & \psi_4^2 & \psi_4^2+\psi_6^2 & & & & \\ 0 & 0 & 0 & -\psi_5^2 & \psi_6^2 & \psi_5^2+\psi_6^2 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_7^2+\psi_8^2 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_7^2 & \psi_7^2+\psi_9^2 & \\ 0 & 0 & 0 & 0 & 0 & 0 & -\psi_8^2 & \psi_9^2 & \psi_8^2+\psi_9^2 \end{pmatrix} \tag{12.18}$$

where we note two things. First, $\breve{\mathbf{\Psi}}^2$ is a symmetric matrix and we have suppressed the upper half. Second, we have displayed $\breve{\mathbf{\Psi}}^2$ in a smaller font for sake of being able to display the entire matrix.

For completeness, we present, in as much as possible, the same computations for the case mentioned earlier with a four-item, forced choice instrument with 3 blocks. In this case, A is an 18×12 block diagonal matrix with the following block structure

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

It then follows that the matrix $\breve{\mathbf{\Lambda}}$ is an 18×4 matrix with "stacked" blocks of the form

$$\begin{pmatrix} \lambda_i & -\lambda_{i+1} & 0 & 0 \\ \lambda_i & 0 & -\lambda_{i+2} & 0 \\ \lambda_i & 0 & 0 & -\lambda_{i+3} \\ 0 & \lambda_{i+1} & -\lambda_{i+2} & 0 \\ 0 & \lambda_{i+1} & 0 & -\lambda_{i+3} \\ 0 & 0 & \lambda_{i+2} & -\lambda_{i+3} \end{pmatrix}$$

where the subscripts in the λ symbols follow a rule similar to that discussed in the earlier example. To put the scope of all this into a more complete picture, we have the following dimensions for the various elements of the model for the TTI Success Insights Style Insights assessment.

1. There are $p = 24$ blocks of $n = 4$ forced-choice items each measuring $d = 4$ traits;
2. This implies $\tilde{n} = \frac{4(4-1)}{2} = 6$ preference comparisons per block;
3. This, in turn, implies that $\gamma = -A\boldsymbol{\mu}_t$ is a vector of $(24 \cdot 6) \times 1$, or 144×1 thresholds;
4. A is a 144×96 matrix, $\boldsymbol{\Lambda}$ is a 96×4, giving $\breve{\boldsymbol{\Lambda}}$ as a 144×4 matrix;
5. Similarly, $\boldsymbol{\Psi}^2$ is a 96×96 matrix and combined with the dimensions of A, we have that $\breve{\boldsymbol{\Psi}}^2$ is a 144×144 matrix.

At this stage in the discussion of the Thurstonian IRT/SEM modeling approach, [30] goes into an explanation of some more technical details not relevant to the current discussion. The intent in this manual is to outline the approach and to provide results when appropriate, not present a complete discussion of topics that are beyond the scope of this document. The interested reader may consult the section on "Identification of Thurstonian IRT Models for Force-Choice Questionnaires" in [30], among other references. Similarly, we have followed the information in [124] for the implementation and estimation of this model to all TTI Success Insights forced-choice format assessment data.

The approach relies on an MPlus (Muethen & Muethen) SEM code generator provided by the authors of [124] in a Microsoft Excel macro. The output is then analyzed using the R Statistical Package in multiple ways. In many cases the software package Matlab is used to generate graphics and scoring algorithms. Similarly, the software package Maple is also used for generation of graphical output based on the results of the model. The next section describes some of the more important output statistics and graphical displays that may be generated based on such output.

12.2 Item Characteristic Functions

We again refer the reader to [30] for a more detailed discussion of the subsequent topics presented here. The main idea at play is that since the latent traits ($\boldsymbol{\eta}$ from earlier) and the unique factors ($\boldsymbol{\epsilon}$ from earlier) are assumed normally distributed, we must have then that the response $\boldsymbol{y}^*$ is also normally distributed. Hence, for each index $l = (i, k)$ we represent the item characteristic function of any paired comparison as a normal ogive model. In other words, the conditional probability that $y_l = 1$, i.e., the conditional probability of preferring item i over item k, conditioned on $\boldsymbol{\eta}$, is given by

$$Pr\left(y_l = 1|\boldsymbol{\eta}\right) = \Phi\left(\frac{-\gamma_l + \breve{\boldsymbol{\lambda}}_l'\boldsymbol{\eta}}{\sqrt{\breve{\psi}_l^2}}\right), \tag{12.19}$$

where we have the following:d

1. $\Phi(\cdot)$ denotes the cumulative standard normal distribution;
2. γ_l is the threshold for binary outcome y_l;
3. $\breve{\boldsymbol{\lambda}}_l'$ is the $1 \times d$ vector of factor loadings;
4. $\psi_l^2 = \psi_i^2 + \psi_k^2$ is the uniqueness of the latent response y_l^*.

Based on the assumption that each item measures only a single trait, we may then consider the item characteristic function (ICF) to be a function based only on comparing item i measuring trait η_a to item k measuring η_b. We may thus simplify the result of (12.19) to

$$Pr\left(y_l = 1 | \eta_a, \eta_b\right) = \Phi\left(\frac{-\gamma_l + \lambda_i \eta_a - \lambda_k \eta_b}{\sqrt{\psi_i^2 + \psi_k^2}}\right). \tag{12.20}$$

The estimation process described in [124] provides estimations of all parameters in (12.20) and we thus have a relatively simple version of the item characteristic function for the comparison of any two items (in a given frame) for a forced-choice assessment format in the form of a standard normal distribution (as a function of the two latent variables, η_a and η_b). This allows the researcher to display the item characteristic functions as 2-dimensional surfaces embedded in $\mathbb{R}^3$.

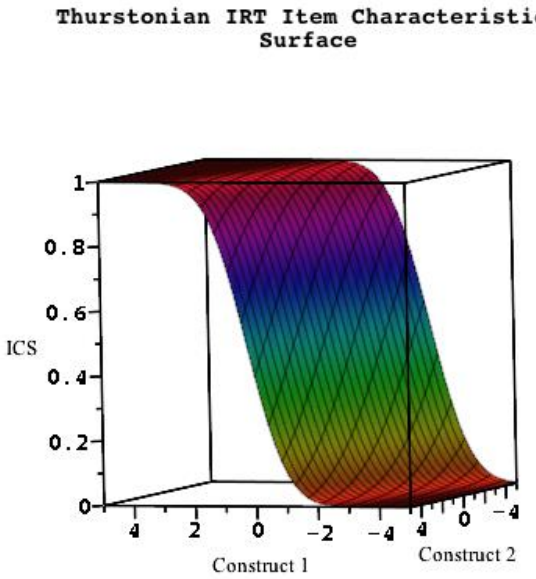

Figure 12.1: Generic Item Characteristic Surface for Thurstonian IRT Model

12.3 Item and Test Information Functions

Figure 12.2 shows the plot of a generic item information surface. The item information function for the Thurstonian IRT model is discussed, for example, in any of [29], [30], [31], [32], [122], and [123]. It is the intention of the authors of this manuscript to include a more in-depth discussion of the item information functions and test information functions in a fashion similar to that presented in Section 11.4. For now, the interested reader may consult any of the references mentioned in the introduction to this section.

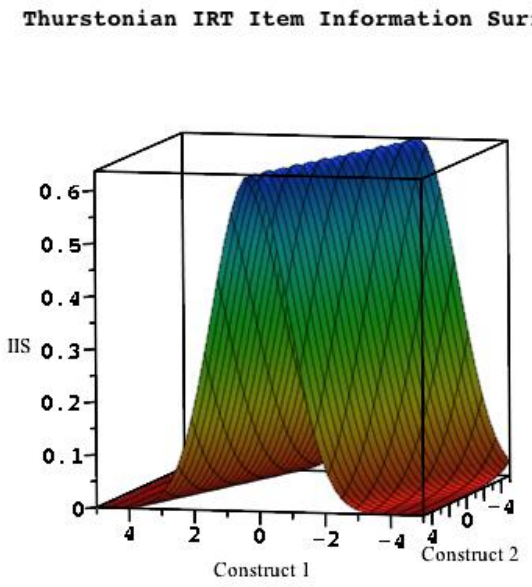

Figure 12.2: Generic Item Information Surface for Thurstonian IRT Model

13 Relationships to Other Variables Evidence of Validity

The American Psychological Association (APA) identifies five key areas of any psychometric assessment requiring evidence of validity, see [156]. One of these areas is *Relationship to Other Variables* which encompasses several concepts such as criterion, concurrent, and predictive validity. While [74] presents many of the theoretical concepts of evidence of reliability and validity of assessments, [5] 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 [5] 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 that 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 Talent Insights® 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 3.3 above for a more detailed discussion presenting some of the reasons for our choice.

In a previous study, [70] shows that the TTI 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 database. This study is an example demonstrating the relationship between TTI assessment derived variables (Behaviors, Motivators, and DNA) with other variables (membership in a specific population within a larger population).

In April, 2018, TTI began offering the opportunity for our network of Value Added Associates (VAA) to add demographics collection to their assessment links. Since that time a good deal 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 [70], 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 demographics database.

This paper is organized as follows. The next two sections discuss some details of the data sets under consideration including an internal consistency presentation. This is followed by a section briefly discussing some of the details of logistic regression. A section discussing measurement of the success or failure of the model is presented next, followed by a section presenting the results of the study on several data sets. Finally, a summary and future work section wraps up the paper.

13.1 Talent Insights Data and the Appropriate Subsets

The TTI Success Insights Talent Insights assessment is comprised of two separate assessments. The first, Style Insights, is used to measure observable behavior and is based on the four factor DISC model. The four factors stand for **D**ominance, **I**nfluence, **S**teadiness, and **C**ompliance. The second

is called Motivation Insights and is based on a six factor model and is used to measure the inherent motivation behind an individual's actions. The scales measured are Theoretical, Utilitarian, Aesthetic, Social, Individualistic, and Traditional.

An additional set of variables is considered during the analysis below. There are a set of 12 variables derived from the DISC variables of the Style Insights assessment. They are jointly called the Behavioral Hierarchy. This set contains the variables Urgency, Frequent Interaction with Others, Organized Workplace, Analysis of Data, Competitiveness, Versatility, People Oriented, Frequent Change, Customer Relations, Follow Up and Follow Through, and Following Policy.

Additionally, Motivation Insights generates a total of 12 separate scores that are collectively known as Driving Forces. In the section containing the results below, the regressions consider the 12 Driving Forces® variables while analysis of the scales focuses on the 6 scales mentioned above. There is no inconsistency in this approach. All information is generated by the same assessment taker responses.

All data considered in this study has been collected between April 2, 2018 and September 14, 2018. The data sets are pulled from the TTI Internet Delivery Service® databases. A data set is generated for each of the two assessments, along with a combined data set generated by individuals taking the Talent Insights combined assessment.

Each of the three data sets comes with a series of demographic data categories that may or may not be utilized in any given instance. For those data entries that contain non-empty demographic information, a brief study of categorical breakdowns reveals a series of subsets of the larger data sets that may be individually studied. Of primary interest for this study is the Job Description category which contains a list of job descriptions as defined by O*NET. For more information on O*NET one may consult the United States Department of Labor Employment and Training Administration O*NET website at https://www.doleta.gov/programs/onet/eta_default.cfm.

Several job classifications are clearly chosen more frequently than others making them particularly important to TTI Success Insights given their higher frequency in the current data sets.

Table 13.1: Most Commonly Chosen Job Descriptions: Talent Insights

Title	Frequency
Sales Rep, Services, All Other	579
Accountants	270
Managers, All Other	265
Sales and Related, All Other	231
CEO	208

The highest frequency job descriptions chosen in the combined Talent Insights data set are summarized in Table 13.1.

Ideally, it is preferable to have larger subsets of data to study. However, it is the opinion of these authors that the frequencies presented in Table 13.1 are large enough to present an initial look at these categories with an eye toward the future when much more data becomes available. The purpose of this initial look is to establish relationships between TTI Success Insights assessment variables and external variables, specifically Job Description. Future studies are planned for such a time when more data are available and, preferably, access to other criteria such as a performance metric are also available.

The data sets used in the study are then formed by taking the set of information of interest and combining them with a randomly generated sample of an appropriate size. For this study, each set of interest is combined with a random sample of size 1,000. No selection criteria are used in addition to the requirement that the random subset comes from the larger Talent Insights data set and that the random subset does not contain any individuals from the population of interest.

We shall use the naming convention as outlined in Table 13.2.

Table 13.2: Naming Convention

Title	Data Set
Sales Rep, Services, All Other	Sales1
Accountants	Accountants
Managers, All Other	Managers
Sales and Related, All Other	Sales2
CEO	CEO

Before discussing some of the statistics related to measures of evidence of reliability and validity, we present some basic information obtained from the demographics for completeness.

Table 13.3: Basic Information Sales1

Category	N	Avg Age	M/F Ratio
Sales1	579	42	398/181
Random	1000	43	555/445

Table 13.3 presents the Sales1 category and shows a set weighted toward males with approximately 69% of the assessment data coming from males. Additionally, the Ethnicity breakdown shows 472 Caucasians with 28 African Americans coming in as the second largest group in this data set. In the Random data set, 767 respondents chose Caucasian with African Americans coming in as second largest at 66.

Table 13.4: Basic Information Accountants

Category	N	Avg Age	M/F Ratio
Account	270	44	119/151
Random	1000	43	538/462

Table 13.4 presents the Accountant category and shows a set weighted toward females with approximately 56% of the assessment data coming from women. Additionally, the Ethnicity breakdown shows 208 Caucasians with 15 Hispanic or Latino coming in as the second largest group in this data set. In the Random data set, 803 respondents chose Caucasian with African Americans coming in as second largest at 55.

Table 13.5 presents the Manager category and shows a set weighted toward males with approximately 60% of the assessment data coming from men. Additionally, the Ethnicity breakdown shows 203 Caucasians with 20 Hispanic or Latino coming in as the second largest group in this data set. In the Random data set, 774 respondents chose Caucasian with African Americans coming in as second largest at 66.

Table 13.5: Basic Information Manager

Category	N	Avg Age	M/F Ratio
Manager	265	45	160/105
Random	1000	42	555/445

As is seen in Table 13.6, the CEO category is quite heavily weighted toward males with approximately 77% of the assessment data coming from males. Additionally, the Ethnicity breakdown shows 184 Caucasians with 11 African Americans coming in as the second largest group in this data set. In the Random data set, 767 respondents chose Caucasian with African Americans coming in as second largest at 69.

Table 13.6: Basic Information CEO

Category	N	Avg Age	M/F Ratio
CEO	208	49	161/47
Random	1000	43	556/444

As is seen in Table 13.7, the Sales2 category is heavily weighted toward males with approximately 63% of the assessment data coming from males. Additionally, the Ethnicity breakdown shows 191 Caucasians with 12 African Americans coming in as the second largest group in this data set. In the Random data set, 784 respondents chose Caucasian with African Americans coming in as second largest at 66.

Table 13.7: Basic Information Sales2

Category	N	Avg Age	M/F Ratio
Sales2	231	44	146/85
Random	1000	43	563/437

13.2 Internal Consistency Estimates for Talent Insights Data Subsets

Internal consistency estimates help assess the consistency of the individual scales of an assessment. There are several measures one may use to assess internal consistency. For purposes of this exposition, the authors choose to present the most commonly reported measure, the α coefficient.

The tables below also present two measures related to evidence of reliability and validity. The first is the average inter-item correlation, which is directly proportional to the value of the α coefficient. The second is the average corrected item-total correlation. Corrected item-total correlation is a measure that may be interpreted to some extent as a proxy for internal structure evidence of validity, see [136].

While there is not a consensus on what specifically are acceptable levels of average inter-item correlation and average corrected item total correlation, we cite [39] for a source on ranges to guide the discussion. For inter item correlation, the authors of [39] state two generally acceptable ranges that depend on the intended scope of the scale. The first range is 0.20 to 0.40 for a narrower scope

and 0.15 to 0.50 for more encompassing scale. Generally speaking, the desired range for corrected item-total correlation is 0.30 to 0.50. The reader should note that these ranges are guidelines, not hard cutoff points.

Whether a measure is acceptable or not is often a matter of interpretation of the scale, its intended purpose, and many other variables. A general rule of thumb is that items with very low item correlation provide little information and those with too high become redundant. The respective averages provide a summation of the overall effectiveness of the scale.

Table 13.8: Common Reliability Coefficient Interpretation, see [59]

Range	Interpretation
$\alpha \geq 0.90$	Shorten 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

Finally, there is a more sound foundation for interpreting the internal consistency measure, although there are differing opinions in this arena as well. However, for the purposes of this study, the authors use the guidelines presented in Table 13.8.

Table 13.9: Style Insights Statistics: Sales1

Scale	Std α	Avg IIC	Avg ITC
D	0.88	0.24	0.46
I	0.85	0.19	0.40
S	0.87	0.21	0.43
C	0.83	0.17	0.37

Table 13.10: Motivation Insights Statistics: Sales1

Scale	Std α	Avg IIC	Avg ITC
The	0.84	0.31	0.51
Uti	0.82	0.27	0.47
Aes	0.83	0.28	0.48
Soc	0.87	0.37	0.56
Ind	0.84	0.30	0.50
Tra	0.80	0.26	0.45

Table 13.9 shows the Style Insights internal consistency and correlations for the Sales1 data set. The computations are performed on the target population. All internal consistency measures score in the Very Good range. The average inter-item correlations are all acceptable according to the larger

range interpretation of 0.15 to 0.50, and the corrected item-total correlations fall nicely within a good range. Table 13.10 displays the same information as Table 13.9 on the Motivation Insights assessment data. In this case all data falls nicely within acceptable ranges for all three statistics presented.

Table 13.11: Style Insights Statistics: Accountant

Scale	Std α	Avg IIC	Avg ITC
D	0.90	0.28	0.50
I	0.85	0.19	0.41
S	0.87	0.21	0.43
C	0.81	0.15	0.36

Table 13.12: Motivation Insights Statistics: Accountant

Scale	Std α	Avg IIC	Avg ITC
The	0.89	0.40	0.60
Uti	0.81	0.26	0.46
Aes	0.82	0.27	0.47
Soc	0.89	0.40	0.59
Ind	0.84	0.30	0.50
Tra	0.83	0.29	0.49

Tables 13.11 and 13.12 present the results for the Accountant data set. All variables from the Style Insights show an acceptable level of consistency and correlation. The possible lone exception is the C variable shows a borderline average inter item correlation. The Motivation Insights variables show a solid performance on all metrics at hand, although two variables show corrected item total correlation that may be slightly on the high side.

The data in Table 13.13 shows an average inter item correlation for the C scale in the Managers data set for the Style Insights variables that is borderline low.

Table 13.13: Style Insights Statistics: Managers

Scale	Std α	Avg IIC	Avg ITC
D	0.87	0.24	0.44
I	0.85	0.19	0.41
S	0.84	0.18	0.39
C	0.83	0.15	0.36

The remainder of the variables show solid scores for all variables and the C variable is in the Very Good range for the internal consistency measure. Table 13.14 shows good average correlations scores for both inter-item and corrected item-total correlation. Four of the six scales fall just under the Very Good range, all at the 0.79 level on the internal consistency metric.

Table 13.14: Motivation Insights Statistics: Managers

Scale	Std α	Avg IIC	Avg ITC
The	0.79	0.24	0.44
Uti	0.79	0.24	0.44
Aes	0.79	0.24	0.43
Soc	0.88	0.38	0.58
Ind	0.80	0.25	0.44
Tra	0.79	0.24	0.43

Table 13.15: Style Insights Statistics: Sales2

Scale	Std α	Avg IIC	Avg ITC
D	0.89	0.25	0.47
I	0.86	0.21	0.42
S	0.86	0.20	0.41
C	0.84	0.18	0.38

Tables 13.15 and 13.16 present the information for the Sales2 data set. These tables show arguably the most consistent data across all variables and metrics of interest of the data considered.

Table 13.16: Motivation Insights Statistics: Sales2

Scale	Std α	Avg IIC	Avg ITC
The	0.85	0.32	0.52
Uti	0.79	0.24	0.44
Aes	0.84	0.31	0.51
Soc	0.88	0.38	0.57
Ind	0.83	0.29	0.49
Tra	0.82	0.27	0.47

Table 13.17: Style Insights Statistics: CEO

Scale	Std α	Avg IIC	Avg ITC
D	0.89	0.25	0.47
I	0.88	0.23	0.45
S	0.84	0.18	0.39
C	0.85	0.19	0.40

The final data set containing the data on Chief Executives is the other data set in the conversation for most consistently across all variables and metrics. The data from both data sets and all metrics considered show the highest levels of inter-item correlation for the D and I scales along with solid consistency across all other metrics for these variables and all metrics for the remaining variables.

Table 13.18: Motivation Insights Statistics: CEO

Scale	Std α	Avg IIC	Avg ITC
The	0.87	0.37	0.56
Uti	0.83	0.29	0.50
Aes	0.78	0.23	0.42
Soc	0.89	0.40	0.60
Ind	0.76	0.21	0.40
Tra	0.80	0.25	0.44

13.3 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 the logistic regression approach.

Generally speaking, there are a multitude of excellent references on logistic regression. This paper follows the work in [89], 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

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{13.1}$$

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 [89], there are two main reasons for choosing the logistic distribution in (13.1). 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.

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right). \tag{13.2}$$

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 [89], 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

$$\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}. \tag{13.3}$$

Note that (13.3) 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 (13.3):

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i} \tag{13.4}$$

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, (13.4) 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 (13.4) 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,

$$\begin{aligned}\mathcal{L}(\beta) = \sum_{i=1}^{n}\{&y_i \ln(\pi(x_i)) \\ &+ (1-y_i)(\ln(1-\pi(x_i)))\}.\end{aligned} \tag{13.5}$$

defines the log-likelihood function.

The problem at hand is now to optimize (13.5) with respect to the parameters $\boldsymbol{\beta}$. It should be noted that while the parameters $\boldsymbol{\beta}$ are not explicitly present in (13.5), one may substitute the definition of $\pi(x)$ from (13.1) into (13.5) to see that (13.5) 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:

$$P(y=0|x) = \frac{1}{1+\exp(g_1(x))+\exp(g_2(x))}, \tag{13.6}$$

$$P(y=1|x) = \frac{\exp(g_1(x))}{1+\exp(g_1(x))+\exp(g_2(x))}, \tag{13.7}$$

and

$$P(y=2|x) = \frac{\exp(g_2(x))}{1+\exp(g_1(x))+\exp(g_2(x))}, \tag{13.8}$$

where

$$g_i(x) = \beta_{i0} + \beta_{i1}x_1 + \ldots \beta_{in}x_n. \tag{13.9}$$

In (13.9) 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 (13.5) 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.

13.4 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 [66]. For some critiques of the approach with additional options offered, one may be interested in [138].

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. 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 $\mathbf{p}$ and $\mathbf{n}$ denote the predicted outcomes and let $\mathbf{Y}$ and $\mathbf{N}$ 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 13.1 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 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

		Prediction outcome	
		p	**n**
actual value	**Y**	True Positive	False Positive
	N	False Negative	True Negative
	total	P'	N'

Figure 13.1: Confusion Matrix

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

$$TPR = \frac{TP}{P'}. \tag{13.10}$$

Similarly, the false positive rate is the total number of false positives divided by the total number of negatives.

$$FPR = \frac{FP}{N'}. \tag{13.11}$$

Some quantities of interest may be accuracy, the total correct classifications divided by the sample size:

$$Acc = \frac{TP + TN}{P' + N'}, \tag{13.12}$$

precision, the true positives divided by the total number of positives, true and false:

$$Pre = \frac{TP}{TP + FP}, \tag{13.13}$$

and a measure of overall fit called the F-measure

$$F = \frac{2}{\frac{1}{Pre} + \frac{1}{Acc}}. \tag{13.14}$$

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:

We may compute the above mentioned statistics.

$$TPR = \frac{17}{25} = 0.68, \tag{13.15}$$

$$130$$

Table 13.19: Example Confusion Matrix

17	12
8	63

$$FPR = \frac{12}{75} = 0.16, \tag{13.16}$$

$$Acc = \frac{17 + 63}{100} = 0.80, \tag{13.17}$$

$$Pre = \frac{17}{17 + 12} = 0.59, \tag{13.18}$$

and

$$F = \frac{2}{\frac{1}{.59} + \frac{1}{.8}} = .68. \tag{13.19}$$

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 I gave were 0.25 and 0.50 for the fictitious data in Table 13.19. 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.

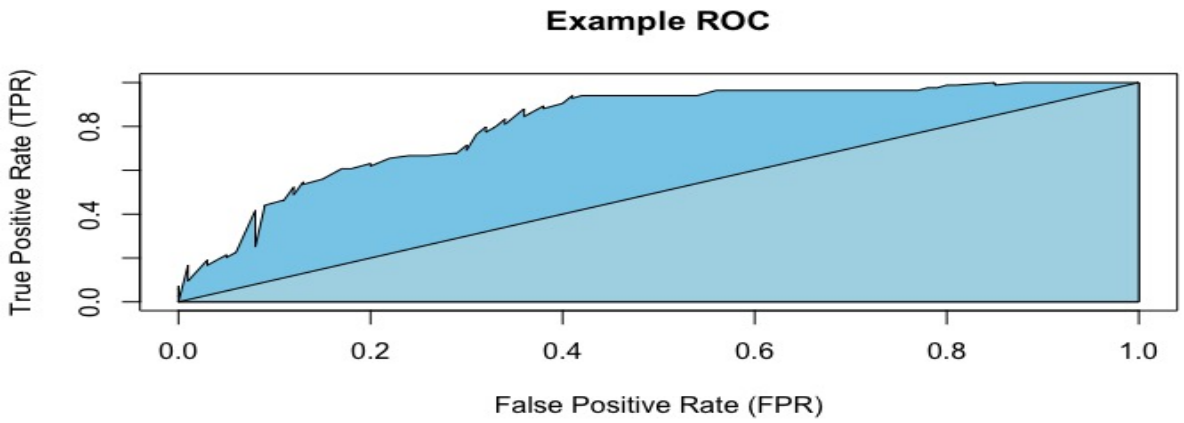

Figure 13.2: Example ROC Curve

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 13.2

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 13.2 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,

$$AUC = \int_0^1 f(x)dx \tag{13.20}$$

where $f(x)$ is a functional representation of the FPR-TPR relationship.

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 [66].

13.5 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 [60].

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

$$cor(X, Y) = -\frac{1}{r} \tag{13.21}$$

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 as much as is possible to avoid over fitting. 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

$$PC_1 = \sum_{i=1}^{n} w_{(1)i} X_i \tag{13.22}$$

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

$$\sum_{i=1}^{n} w_{(1)i} = 1. \tag{13.23}$$

We then compute the second new vector PC_2 which is now 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 orthogonal 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.

13.6 Job Description Classification Results

This section presents some of the results of the current study. Any results not presented in this section are presented in full in Appendix **A** below. For brevity of exposition, results of this section are based on the CEO sample unless otherwise explicitly stated.

We begin with an application of the logit transformation presented in 13.2 as a tool to help determine whether variables of interest are candidates for inclusion in a logistic regression model.

Figure 13.3 presents a good example of a linear fit of the log odds of membership in an interval of a score on a scale, a desired property for inclusion. This particular case shows the Natural Dominance scale from the CEO data set from the Style Insights assessment. The log odds are computed using (13.2) where $\pi(x)$ is replaced by p and p represents the conditional probability of having a score in the interval of interest given membership in the group of interest.

The data shown in Figure 13.4, in contrast to the results of the Natural Dominance graph in Figure 13.3, shows no linear relationship at all. This data represents the Intentional scale from the Motivation Insights portion of the CEO data. The takeaway is that the Natural Dominance scale is a reasonable candidate for direct inclusion in a logistic regression model for the CEO data while the Intentional data is not.

The linear relationship of the log odds of the data to membership in a group and bucket combination determines both the relative potential strength of the variable as a predictor and the directionality expected for any coefficient related to that variable in any logistic regression approach. As an example, one would expect that the Natural Dominance variable has a positive coefficient given its strong positive linear relationship. A failure to maintain this directionally correct relationship in the coefficients could be cause for removal of a variable from consideration.

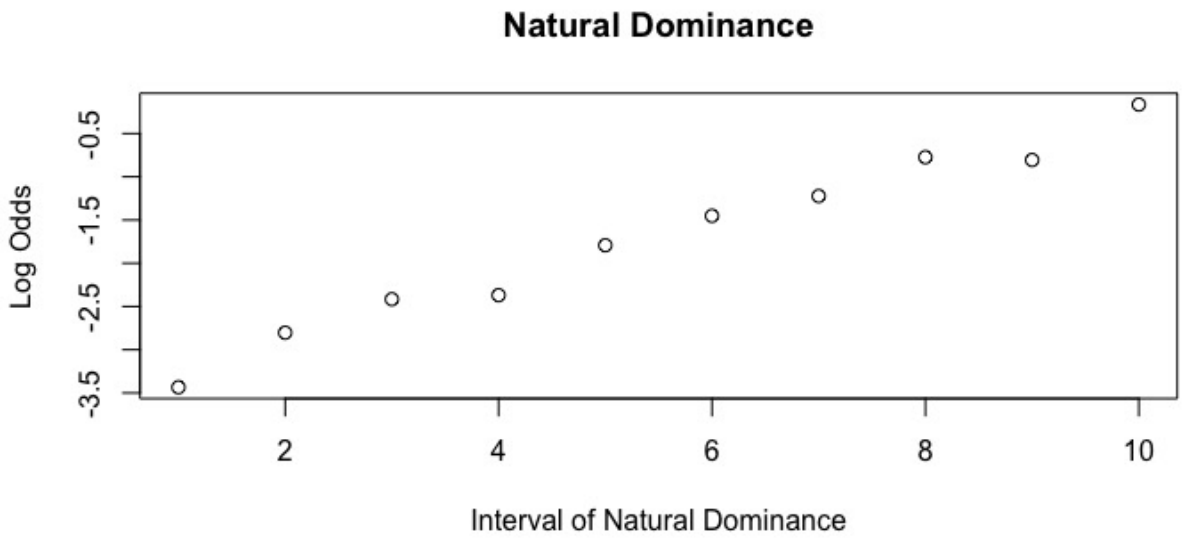

Figure 13.3: Natural Dominance Log Odds Plot

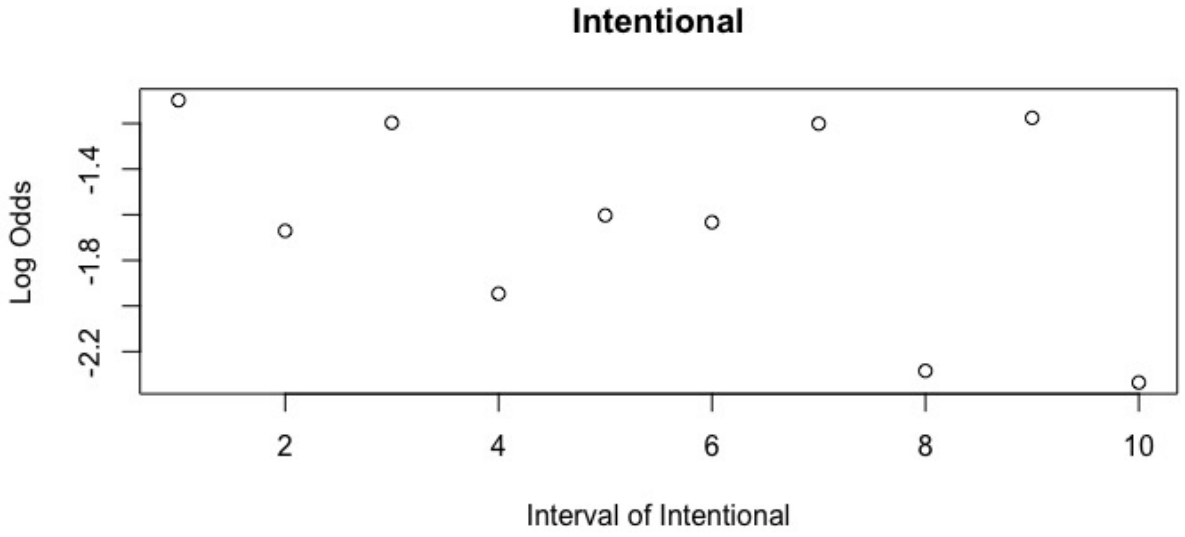

Figure 13.4: Intentional Log Odds Plot

As noted in a previous section, the ROC analysis is concerned with determining the relationship between the true positive rate (TPR) and the false positive rate (FPR). Further, we wish to determine whether the information taken from the assessment variables generates identification rates better than a random sampling technique.

This may be done in at least two ways. The first is to visually inspect the ROC curve and its relationship to the line $y = x$ in the plane. The second is to compute the area under the ROC

curve and compare that to the area under the curve generated by the line $y = x$ on the interval $[0, 1]$. Recall that this area is equal to 0.50. Examples of both methods are presented in a previous section of this paper. Additionally, many statistics are available to be compared across data sets and to help determine additional information about the quality of the fit. Given that this paper is interested in determining whether relationships exist between variables of interest, as opposed to attempting to generate a fully predictive model and testing its accuracy, the authors choose to present the Brier statistic, see, e.g., [138].

First presented is the area under the curve (AUC) statistic. Table B.141 shows the results of the analysis applied to the CEO data set. The results are fairly consistent across the data subsets with Behaviors and All DISC (Behaviors with Behavioral Hierarchy included) scoring nearly identically, and showing a slight advantage over the scoring of the Driving Forces data subset. When considering the full Talent Insights data set, we see an increase in AUC of nearly 9% (0.75/0.69-1).

Table 13.20: CEO Sample
Area Under Curve

Subset Considered	No PCA	PCA
Behaviors	0.71	0.71
All DISC	0.71	0.71
12 Driving Forces	0.70	0.69
All Variables	0.74	0.75

The Brier statistics show very little in the way of differences between the models, with only a slight change in the all variables version. The No PCA version loses a 0.01 to the earlier versions, while the PCA version gains a 0.01 advantage.

Table 13.21: CEO Sample
Brier Statistic

Subset Considered	No PCA	PCA
Behaviors	0.13	0.13
All DISC	0.13	0.13
12 Driving Forces	0.13	0.13
All Variables	0.14	0.12

Figure 13.5 presents the ROC curve compared to the random selection technique. The curve is smooth throughout and clearly lies above the line $y = x$ consistent with the values of AUC being consistently above 0.50 as shown in Table B.141.

The authors would like to draw a comparison, again, to an earlier, similar study conducted on Serial Entrepreneurs. In this case, we draw the comparison between % Correctly Classified as presented in [70] and the AUC values in Table B.141. This comparison is presented in Table 13.22. Note that the column Correct ID is a real number version of the % Correctly Classified as reported in the original work.

There are a few things to make note of. First, there is no generalization between identification problems based solely on the variables present. In fact, [70] considered data from a TriMetrix

Legacy DNA assessment, collected in 2010, which looks at the six motivation scales compared to the 12 Driving Forces considered in this work, for which data was collected between April and Spetember, 2018. Similarly, this work does not differentiate between Adapted and Natural Behaviors while [70] considers them separately. The AUC scores presented in Table 13.22 are based on the PCA column of Table B.141.

Table 13.22: CEO v Entrepreneur Comparison

Subset	AUC	Correct ID
All DISC	0.71	0.71
12 Driving Forces	0.69	0.67
All Variables	0.75	0.76

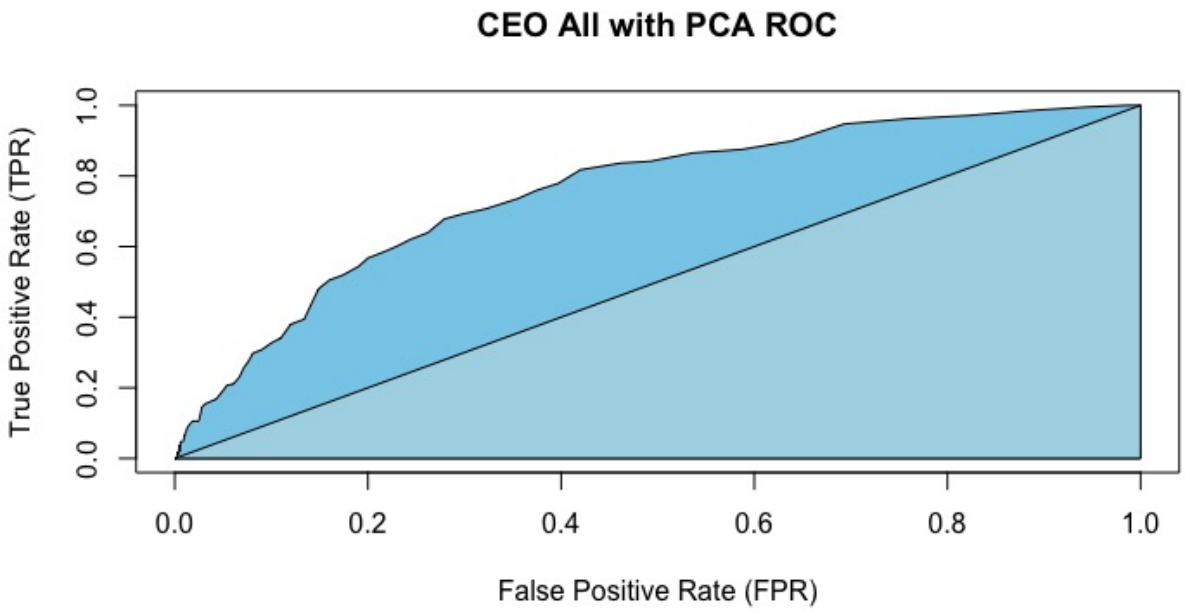

Figure 13.5: CEO ROC Curve: All Variables

What the authors found striking were the similarities between the performance of a very similar modeling approach (logistic regression) on two separate and assumed distinct data sets. To elaborate, both groups of individuals are self-identified as Serial Entrepreneurs or CEO. Serial Entrepreneur is defined as one who has started more than one business with no reference to the success or failure of any business started by the individual. CEO is an abbreviation of a job description in the O*Net classification and is used to denote Chief Executive.

While one may expect some similarities to exist, there is by no means any requirement for any specific qualities to be adhered to by any individual starting a business or by anyone advancing through the ranks of a company to attain a C-level position.

When comparing the results presented in Table 13.22, it is interesting to see that the different subsets considered perform so similarly. Additionally, the hierarchy is generally respected in that each individual subcomponent of Talent Insights performs approximately the same across the subsets and across the two different studies, and combining the data together provides a better result. This final piece of information is critical.

A conclusion of [70] is that additional assessments provide additional information. The results of the current study appear to support that conclusion and support continued evidence of the relationship between Talent Insights variables and external variables, in this case Chief Executives as self-reported and based on the O*Net job classification model.

As a final note, one may question the comparison of the AUC to the % Correctly Identified. To address this, note the % Correctly Classified, as presented in [70], is based on the Confusion Matrix as defined earlier in this paper. For any specific value $x \in [0,1]$ and a probability model, one may define a Confusion Matrix. For the results in [70], the assumption was made that the ratio of membership to non-membership in the set of interest was known and used to generate the Confusion Matrix (50%).

The ROC analysis presented in the current study are based on using a discrete approximation of the ROC curve, which is the functional relationship between the true positive rate and the false positive rate. The AUC statistic is based on a discrete approximation of the ROC curve and generates an approximation of the area under the true ROC curve. To be explicit, denote by $f : \mathbb{R} \mapsto \mathbb{R}$ any continuous function defined on the real line. Then,

$$\frac{1}{b-a}\int_a^b f(x)dx \tag{13.24}$$

defines the average value of f over the interval $[a,b] \subseteq \mathbb{R}$. If we let f be a function representing an ROC curve, we then have

$$AUC = \frac{1}{1-0}\int_0^1 f(x)dx = \int_0^1 f(x)dx. \tag{13.25}$$

Given that the % Correctly Identified from [70] is simply a point on the ROC curve, we can conclude that the AUC values discussed in this paper and the classification percentages from [70] may be compared as presented above.

13.7 Summary and Future Work

This paper has shown that the use of a logistic regression algorithm as an identification technique successfully outperforms the standard random selection technique it is measured against in a statistically significant way, see Section 13.8 for the statistical analysis results. In all five data sets we see several sound points to take away from this work.

First, in all cases, all data sets outperform the random selection technique to varying degrees. The CEO, Accountant, and Sales1 data sets show the strongest differentiation from the random selection approach. The Manager and Sales data sets do not have the strength of differentiation of the first three, yet still outperform random selection.

Second, in all data sets considered, the use of the full set of Talent Insights variables, i.e., both DISC and 12 Driving Forces, produces the best identification. In most cases this approach, combined

with the PCA variable reduction technique, shows strong improvement over using either DISC or 12 Driving Forces variables alone.

Third, the results in this paper confirm earlier results presented in [70] and highlight that the Talent Insights assessment may be successfully and statistically significantly used to classify the groups mentioned in the earlier study as well as the five new groups considered in the current work.

Future studies are scheduled to begin as data becomes available. As demographic data collection grows, the number and scope of studies such as this one are anticipated to grow dramatically.

13.8 Identification Results

Before presenting the list of completed relationships and their levels of success, we discuss some of the variables that come through in the earlier analysis of the CEO data set. It is interesting to note that it is not always the case the expected variables come through with the greatest statistical significance. The CEO sample is a good example of this. We will not go into a complete analysis of each possible variable, nor will we present an in-depth analysis of the PCA variables. Such discussions are anticipated to come at such time as more data is available for study.

For CEO, one may expect that Dominance, either natural or adapted, would be the main identifier from the DISC variables. This is, in fact, the case as Natural Dominance provides the best statistically significant fit. When considering the Behavioral Hierarchy, one may expect variables such as Competitive to come through, and it does. However, we see Organized Workplace, and surprisingly with a negative coefficient, Follow Up and Follow Through. Similarly, for Driving Forces, one may expect Commanding to come through, yet we see Intellectual, Harmonious, and Structured, all with negative coefficients. In the final model with all variables considered without the PCA approach, the model produces Natural Dominance, Intellectual, and Structured, the last two with negative coefficients.

This is not to say that a CEO is not Commanding or that a CEO does not possess the characteristics of high levels of the Influence or Compliance scales. It means that the model is identifying other variables that are able to help identify this particular group of CEOs from the random sample chosen to compare them against. These are the differentiators in this particular study.

The remainder of this section presents the results of the identification modeling approach applied to the remaining four data sets discussed throughout this work. Tables 13.23 and 13.24 along with Figure 13.6 show the results for the data set called Sales1.

Table 13.23: Sales1 Sample
Area Under Curve

Subset Considered	No PCA	PCA
Behaviors	0.66	0.66
All DISC	0.67	0.66
12 Driving Forces	0.63	0.64
All Variables	0.69	0.70

The AUC statistics for Sales1 are not as strong across all variables considered when compared to the CEO sample. However, the results do show the same pattern as the CEO sample and that of earlier work in [70]. All variables show performance that well exceeds that of random sampling

and combining the variables from both DISC and the 12 Driving Forces outperforms using either assessment individually.

Table 13.24: Sales1 Sample
Brier Statistic

Subset Considered	No PCA	PCA
Behaviors	0.22	0.21
All DISC	0.21	0.21
12 Driving Forces	0.22	0.22
All Variables	0.22	0.21

Similarly, the Brier statistics are all at an acceptable level, although not as strong as the CEO sample. The visualization of the AUC comparison to a random sampling approach is displayed in Figure 13.6. The main variables of interest from DISC are Natural Compliance with a negative coefficient and Versatility from the Behavior Hierarchy. From Driving Forces, we see Selfless with a negative coefficient. The three aforementioned variables form the best fit model for Talent Insights and Sales1.

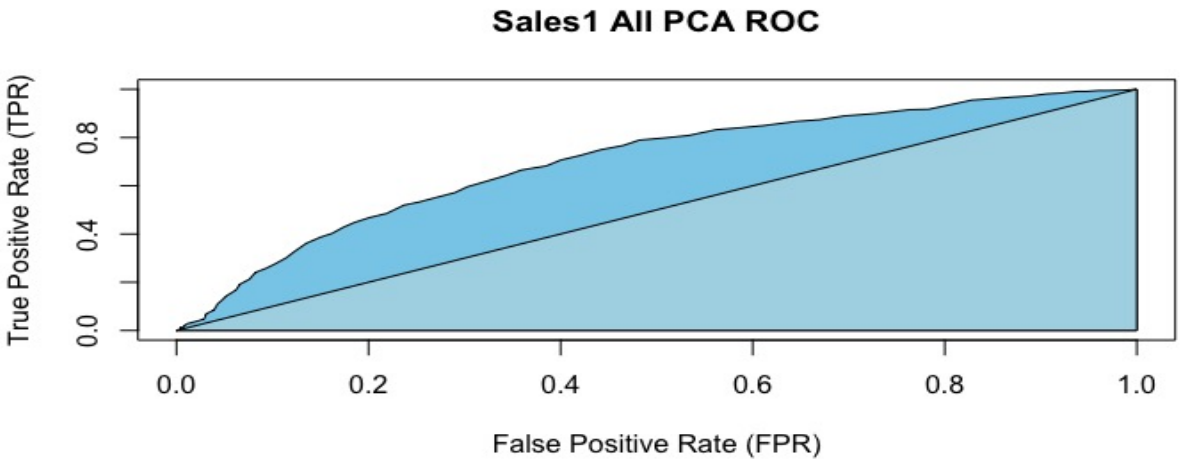

Figure 13.6: Sales1 ROC Curve: All Variables

Tables 13.25 and 13.26 along with Figure 13.7 show the results for the data set called Accountant.

Table 13.25: Accountant Sample
Area Under Curve

Subset Considered	No PCA	PCA
Behaviors	0.67	0.67
All DISC	0.67	0.67
12 Driving Forces	0.59	0.63
All Variables	0.70	0.70

The results in Table 13.25 show that the DISC variables perform similarly to both the Sales1 and CEO samples, while there is some drop off in the performance of the 12 Driving Forces in this sample. However, all samples outperform the random sampling approach, and the combination of DISC and 12 Driving Forces performs best.

Table 13.26: Accountant Sample
Brier Statistic

Subset Considered	No PCA	PCA
Behaviors	0.16	0.16
All DISC	0.16	0.16
12 Driving Forces	0.16	0.16
All Variables	0.17	0.15

The Brier statistics shown in Table 13.26 are similar to those in the CEO sample, showing slightly better results than the Sales1 sample. The AUC representation is shown in Figure 13.7. The Accountant sample has Adapted Steadiness and Compliance along with Following Policy, all strongly positively related to the data. Driving Forces has Intellectual and Commanding, with a negative coefficient, come through. The final model for all Talent Insights with no PCA has Following Policy, Commanding with negative coefficient, and Objective come through as the best fit model.

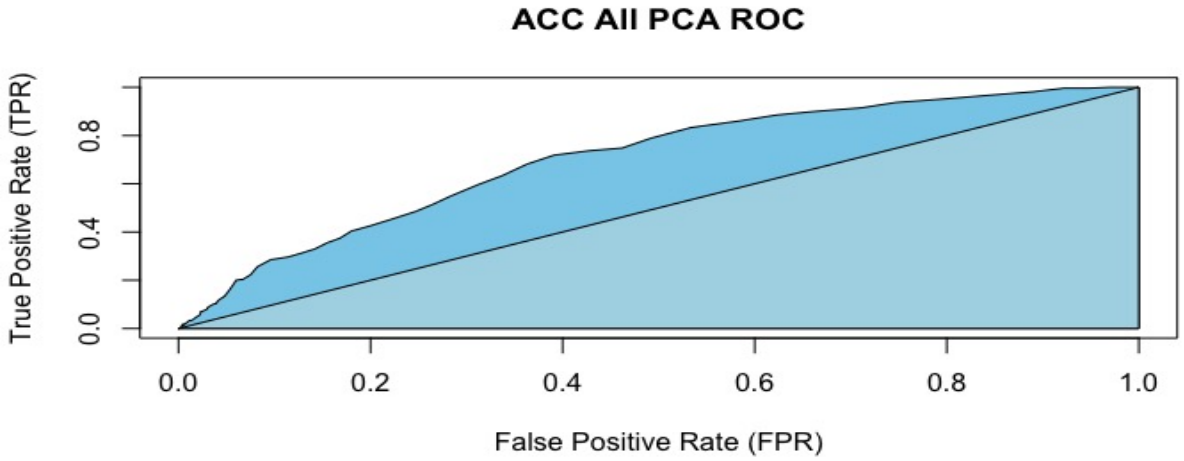

Figure 13.7: Accountant ROC Curve: All Variables

Tables 13.27 and 13.28 along with Figure 13.8 show the results for the data set called Manager. The results in Table 13.27 show that the DISC variables perform similarly to the Serial Entrepreneur version of either Adapter or Natural DISC scales with some drop off in the performance of the 12 Driving Forces in this sample compared to the three earlier samples in this work. However, all samples outperform the random sampling approach and the combination of DISC and 12 Driving Forces performs best once again.

The Brier statistics shown in Table 13.28 are similar to those in the CEO sample, showing slightly better results than the Sales1 sample.

Table 13.27: Manager Sample
Area Under Curve

Subset Considered	No PCA	PCA
Behaviors	0.58	0.58
All DISC	0.58	0.59
12 Driving Forces	0.63	0.59
All Variables	0.63	0.64

Table 13.28: Manager Sample
Brier Statistic

Subset Considered	No PCA	PCA
Behaviors	0.17	0.16
All DISC	0.17	0.16
12 Driving Forces	0.16	0.16
All Variables	0.16	0.16

The AUC representation is shown in Figure 13.8. The results of the regression analysis on this particular data set struggled to identify across the population with all sets of variables. For example, DISC only regression showed Natural Dominance come across as statistically significant. For Driving Forces, Commanding comes through statistically significant. However, there is little improvement shown when combining the data into a Talent Insights combined regression.

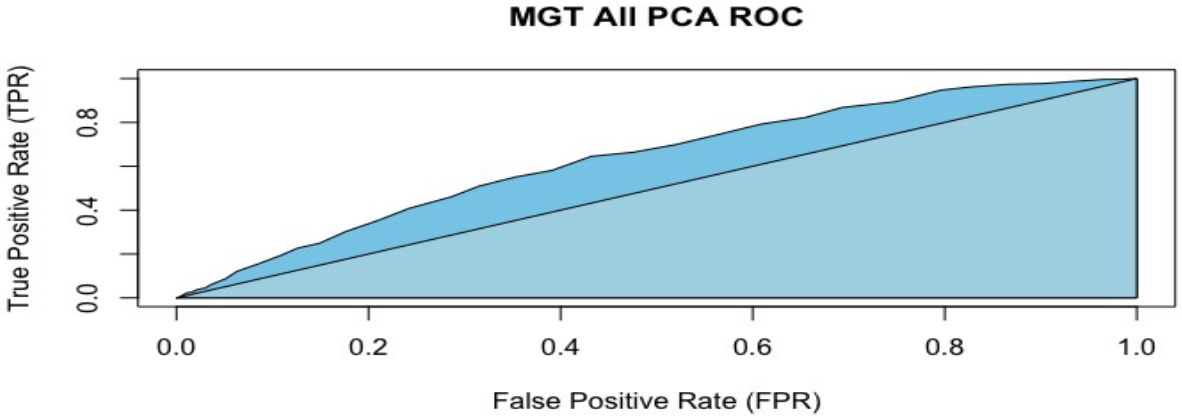

Figure 13.8: Manager ROC Curve: All Variables

When combining all variables in the PCA approach, we see a very limited improvement in the overall output. The conjecture we have is the job classification Manager is quite vague and can run a quite wide spectrum of the population. A manager could be anything from a teen-age assistant manager at a fast food chain restaurant to a middle-aged individual running a convenience store and so on. The possibilities are many, and we believe that this is an obstacle to such a generically defined subset being reasonably identified. Having said that, we still show better than random identification with this model.

Tables 13.29 and 13.30 along with Figure 13.9 show the results for the data set called Sales2.

Table 13.29: Sales2 Sample
Area Under Curve

Subset Considered	No PCA	PCA
Behaviors	0.63	0.62
All DISC	0.63	0.63
12 Driving Forces	0.65	0.64
All Variables	0.65	0.66

The results in Table 13.29 show that the DISC variables in all samples outperform the random sampling approach, and the combination of DISC and 12 Driving Forces performs best.

Table 13.30: Sales2 Sample
Brier Statistic

Subset Considered	No PCA	PCA
Behaviors	0.15	0.15
All DISC	0.15	0.15
12 Driving Forces	0.15	0.15
All Variables	0.19	0.15

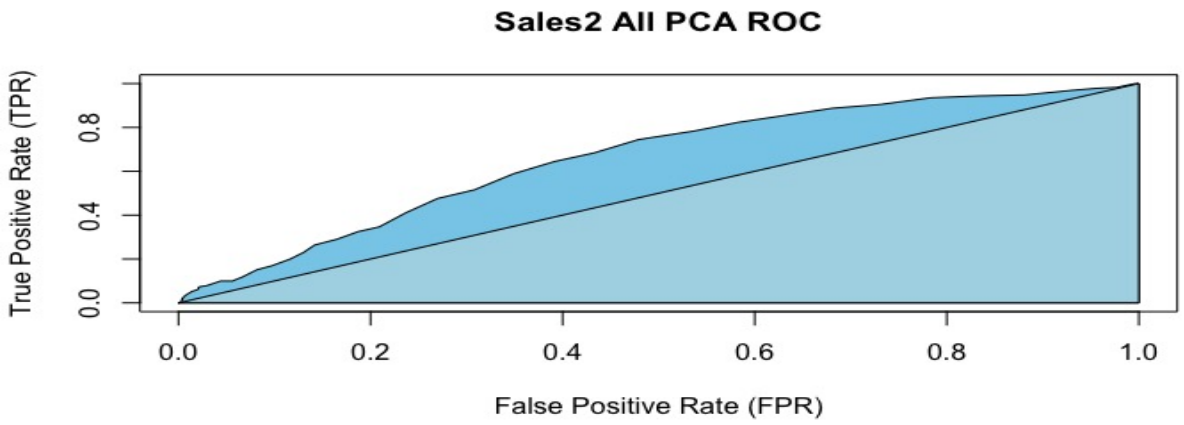

Figure 13.9: Sales2 ROC Curve: All Variables

The Brier statistics shown in Table 13.30 are similar to those in the CEO sample, showing slightly better results than the Sales1 sample with the exception of the All Variables with no PCA category. The AUC representation is shown in Figure 13.9.

The main takeaway from the information presented in this section is that all data sets for all subsets of variables show two main results. First, in all cases, all subsets of variables have statistically significant logistic regression fits that outperform a random sampling technique as shown by the ROC curves and the AUC and Brier statistics. Second, in all cases, the combination of both DISC variables and 12 Driving Forces variables consistently outperforms using either set of assessment variables alone.

In some cases the performance is not dramatically improved. For example, the Sales2 data set shows very similar AUC for 12 Driving Forces and All Variables. The difference is seen in selecting the PCA version over the No PCA version and basing this decision on the improved Brier statistic. In the Accountant, CEO, and Sales1 data sets, the improvement is more noticeable. Similarly, in the Manager data set using the PCA approach shows stronger improvement.

13.9 Additional Results

During the Fall of 2020 the TTI Success Insights Research and Development team conducted a follow on study using only the Style Insights data. Given that the TTI Success Insights demographics collection is aligned with the O*NET job classification system and that Style Insights is the most commonly used assessment in our suite of assessments, it is a natural study to consider. The relevant information and results are included in Appendix B. We only provide a brief description here for convenience.

The follow on study was limited to direct logistic regression only. In other words, the principal component analysis discussed earlier in this section was not considered. The main reason for this is the scope of the study. It is the desire of TTI Success Insights to eventually show evidence of relationships between as many of our assessment variables and the O*NET occupational titles. Given that O*NET has more than 1,000 occupational titles and the requirement of a minimum number of data points in order to ensure stability and statistical significance of the results, the current study considers just over 200 of the aforementioned occupational titles.

A decision was made to consider only those occupational titles that TTI Success Insights had 100 or more respondents in our database. In the event that an occupational title had more than 500 respondents, a decision was made to incorporate a holdout sample testing approach. The holdout sample approach employed randomly created five 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 used a different subset of four of the five data samples as training data. The results of the training model were 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 comes through the modeling process as statistically significant, the relationship may be considered more robust than if they do not.

14 Response Processing Evidence of Validity

The American Psychological Association (APA) Handbook of Testing and Assessment in Psychology, see [74], 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 [156]. 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 [5] 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 [134], 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".

14.1 Using Electroencephalograph (EEG) to Measure 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 [17] and [18]. 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 [21], [22], [23], [26], [27], [46], [47], or [51].

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.

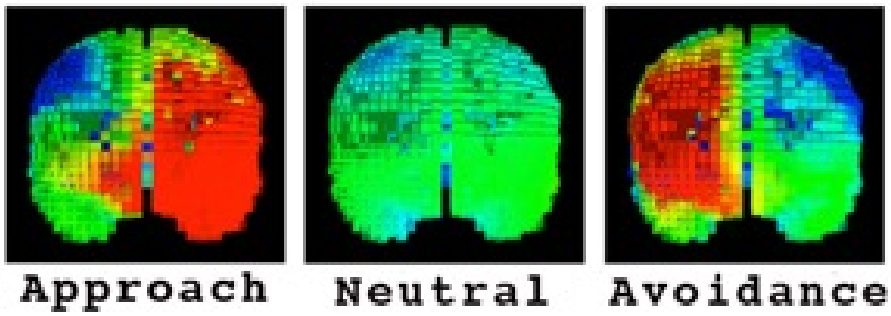

Figure 14.1: An Example of Asymmetry in EEG Captured Brain Activity

Figure 14.1 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.

At the time of publication of this manual, due to the 2020 COVID-19 pandemic, the TTI Success Insights EEG lab was shut down, thus limiting our EEG response data for the Style Insights assessment. However, protocols are in place to measure not only the direction of the brain response, but also the intensity of specific items. To accomplish this, each forced-choice frame is modified into four Likert-style response items for participants to make a selection. The existence of these protocols allow the TTI Success Insights research team to collect such data and make comparisons of the brain findings with the respondent's original Likert-style survey responses, as well as an item-by-item comparison to our mathematical findings.

This same protocol has already been employed on our Emotional Quotient assessment and the pilot results are to be published in a forthcoming article. The study protocols are published in [26]. The initial findings from our EQ EEG study found five emerging groupings of items that when combined with mathematical data are lighting the way to continual improvement. Each study participant responded to a subset of items from the TTI Success Insights EQ assessment on a five-point Likert-style scale. The Likert-style scale is anchored with 1 - Strongly Disagree, 2 - Disagree, 3 - Neither Agree nor Disagree, 4 - Agree, and 5 - Strongly Agree. The images that follow are captured at one-eighth second intervals.

Figures 14.2 through 14.6 provide examples of each of the five groupings. In the first, we find an item that has a match between the Likert-style response and their mental processing, which documents a dominant left frontal lobe gamma activity and a match with their agreement rating. The second example is quite different. While the respondent agrees strongly with the item, their brain is slower to form a response, and, when it does, it shows a right side or aversion to the concept. This suggests that the item may have a socially acceptable answer that is arrived at after contemplation by the brain, but results in a very different answer. Both the answer provided and the brain activity in the third example suggest confusion or a lack of discrimination. Example four is a good example of what happens in the brain when confronted with double negative statements. While assessment designers once thought these mental manipulations kept respondents on their toes, this research suggests they are simply difficult to mentally process. The final example is interesting in that it provides very little brain activity suggesting that the participant is not emotionally associating with the item and, therefore, may not be providing additional factor information.

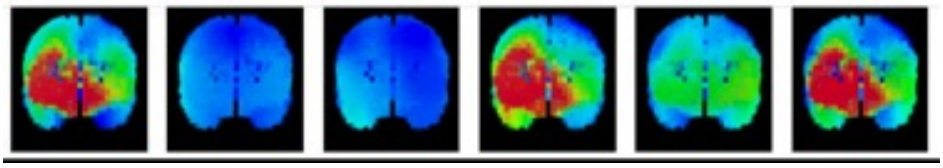

Figure 14.2: Confirmation Between the Survey and Brain Response
Respondent Answer: 4

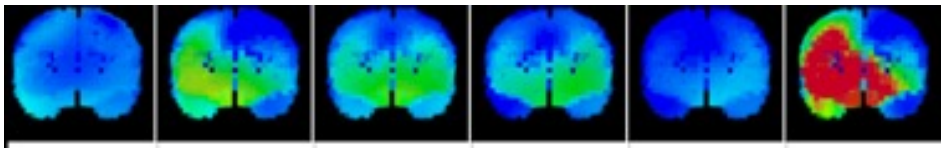

Figure 14.3: Possible Socially Acceptable Response Based on Brain Activity
Respondent Answer: 5

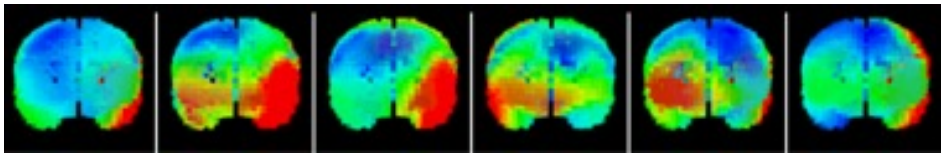

Figure 14.4: Brain Response to Confusing Items
Respondent Answer: 3

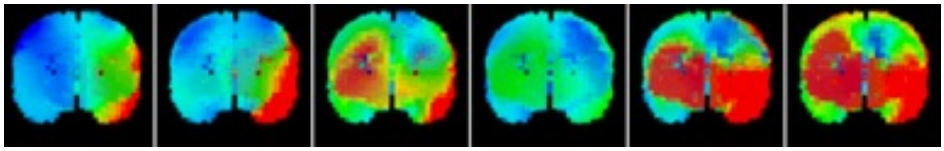

Figure 14.5: Brain Response to Double Negative Items
Respondent Answer: 5

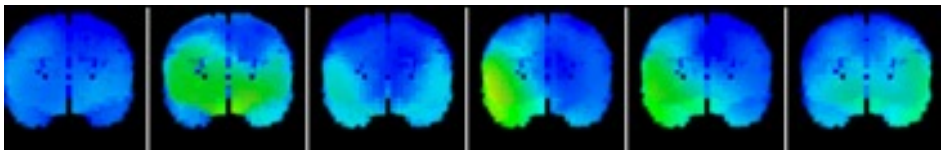

Figure 14.6: Neutral Brain Reaction
Respondent Answer: 5

14.2 Response Processing Summary

While response process evidence of validity continues to be a challenging assessment design hurdle, 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 offers new possible explanations regarding respondent mental processing. We are entering a new era where industrial and organizational assessment 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 decision.

15 Consequences of Testing Evidence of Validity

Consequential validity evidence requires the collection of evidence for both the intended and unintended consequences associated with the use and interpretation of the scores based on an assessment.

There remains a debate over whether testing consequences are extraneous to validity, see [156] pp. 78, or are a critical element in evaluating the appropriateness of using an assessment for a particular purpose, [128]. Even the American Psychological Association (APA) appears to provide somewhat contradictory statements. As just noted, the APA first claims that consequences of testing are extraneous to the establishment of evidence of validity, [156] pp. 78, then co-author [5] which points out the importance of evaluating the appropriateness of a particular use or purpose for an assessment, see Standard 1.25.

The fact remains that virtually all assessment programs have consequences on some level. This fact requires that care be given to evaluate the degree to which the positive outcomes of a test outweigh any potential negative consequences. While this component of validity is acknowledged, it is unclear how much this helps in establishing an overall argument for validity of an assessment. Based on the previous brief discussion, the main question to be asked and answered is whether the test is fulfilling its intended purpose. In order to ask this question, one must first determine the stated purpose of the assessment.

On pp. 429 of [154], the author proposed three questions to help clarify the purpose: "What does the testing practice claim to do? What are the arguments for or against the intended aims of the test? What does the test do in the system other than what it claims, for good or bad?"

15.1 The Intended Purpose of TTI Success Insights' Style Insights Assessment

TTI Success Insights DISC Universal Language Reference Manual states:

> We live in a laboratory where we have the unique opportunity to learn about people. By developing a strong command of the DISC language, you will:
>
> 1. Know your behavioral style;
> 2. Recognize the behavioral style of others;
> 3. Adapt and blend your style for greater, more effective communication and relationships. (pp. XIII)

By exposing "how" individuals interact, Style Insights may also assist employers in the selection process. When hiring agents of an organization understand the unique behavioral characteristics of each job, they can make better hiring decisions. When managers understand the behavioral styles of each employee, they can tailor their management style and re-frame concepts in ways that will better communicate to each employee. Therefore, Style Insights not only allows for greater self-awareness regarding our own behaviors, but also creates opportunities for better job fits.

15.2 Addressing Consequences of Testing

Once the purpose is established, responses that fulfill the intended outcome constitute evidence of positive consequences. It is important to note that [5] clearly states

> Ensuring that unintended consequences are evaluated is the responsibility of those making the decision whether to use a particular test, ... (pp. 31)

As a result of third-party involvement in the form of distributors, addressing negative consequences is a bit more complex. However, TTI Success Insights endeavors to address as many as possible by providing updated adverse impact reports based on measuring possible disparate impact in hiring, see [91], ongoing education, and assessing possible negative consequences that may be attributed to the assessment. These include, but are not limited to, scoring, uses, and interpretations of our assessment scores and more. All of these are identified and mitigation processes established when possible.

As an example, one of the current and most popular uses of the TTI Success Insights Style Insights assessment is in the coaching and professional development space. Generally speaking, TTI Success Insights does not directly participate in the coaching or professional development related to the use of our assessments. Rather, a network of Value Added Associates (VAA) performs this important role. To facilitate the proper use and interpretation of the results of our assessments and their corresponding reports, TTI Success Insights VAA have at their disposal training and certification processes.

TTI Success Insights has established a number of activities and training steps to assist the VAA, including: online training courses, certification programs, and downloadable debriefing guides that walk the user through the report. Recently, TTI Success Insights created an online learning platform, LEARN, specifically designed to guide users through each assessment's intended purpose and appropriate application, https://learn.ttisi.com/login/.

TTI Success Insights also provides an extensive start-up package for new distributors that includes attendance at one of our monthly bootcamps. The three-day training, until recently, was held in person. Due to the COVID 19 pandemic, a decision was made to move the training to the virtual world. One of the keys to the success of bootcamp is the pre-work, which includes foundational assignments that allow the participants to learn and be tested over basic content knowledge prior to attending the three-day workshop. This approach allows for more time to fully address issues, including intended and unintended consequences of assessment score use and interpretation.

In an effort to promote an understanding of the consequences of testing, all VAA and their associates, plus certain industry professionals (VAA clients) are able to take the aforementioned online courses and certification processes. While certification is not currently mandatory, it is recommended that VAA who certify their clients renew their certification bi-annually.

It would be an injustice to our customer support team to not point out their role in monitoring and correcting consequences of testing issues. Each week, this team responds to over 400 calls from assessment distributors around the world. As a result of their direct communication, we are able to help assessment distributors identify the appropriate assessment for their unique situation, as well as addressing any interpretation issues as soon as they arise. It is also worth noting that this line of communication helps connect the TTI Success Insights research team with any issues when they arise in our network.

15.3 Consequences of Testing Summary

While many aspects of testing fall outside the direct control of TTI Success Insights, our continual improvement protocols are in place, and the knowledge of this important standard is part of our improvement processes. TTI Success Insights continues to clarify appropriate assessment use and interpretation along with educating out clients regarding safeguarding against unintended consequences.

16 Summary and Future Work

The main goal of this manuscript is to present a basic outline of the efforts employed by TTI Success Insights to establish evidence of reliability and validity of our Style Insights assessment, as well as reporting the results of such efforts. This section shall serve as a summary of what has been presented elsewhere in this work. Additionally, we discuss some of our thoughts for future projects and studies here at TTI Success Insights.

16.1 Internal Consistency Reliability

Internal consistency reliability is one of the four main areas of reliability as discussed by the American Psychological Association (APA) [74]. It is noted that many authors have discussed the frequency with which the so-called Cronbach's α coefficient is used as a measure of this statistic, see [38] and [86] for example. Cronbach himself suggests that the α coefficient is not the best measure of internal consistency and suggests better measures exist, see [57].

Having just noted Cronbach's own opinion of the α coefficient, we note that we also report this as one of our measures of internal consistency, see Section 7. In the same section we also report our findings based on the factor analytic based McDonald's ω. Finally, we present a discussion of item and test information function approaches to reliability, in the single variable IRT approach, see Section 11.4. Similar results are presented for multivariable item and test information surfaces in Section 12.2.

In the cases of α and ω coefficients, all coefficients for all constructs are in the range of 0.85 to 0.90 with tight confidence intervals. This suggests quite solid internal consistency reliability as based on these measures, with all falling in the "Very Good" range as shown in Table 7.2. Further, we report the α if item deleted statistic, in part as a measure of consistency of the items, but also as a diagnostic tool to determine if any items are significantly negatively impacting any of the scales. Note that historical internal consistency data are presented in Appendices F (internal consistency) and G (α if item deleted).

As discussed at the end of this summary, see 16.11, we currently employ single variable IRT models as diagnostic tools as well as tools for more suitably pairing items in an attempt to improve our forced-rank frames. As continual improvement is a common theme at TTI Success Insights, we use as many tools as possible to provide us with as much information as we are able to obtain. We are also in the beginning stages of implementing a multivariable IRT model that is more theoretically correct for the assessment format.

16.2 Temporal Consistency Reliability

Temporal consistency, or test-retest consistency, is presented in Section 8. Simply put, we wish to determine the consistency of a known subset of respondents over time. This concept is on a relative rather than absolute grounding. In other words, we are not necessarily interested in a single individual scoring different on the two administrations of the assessment. We are interested in the relative standing of all individuals in the sense that there is a strong correlation between the two sets of assessment scores (on a scale by scale basis).

Our most recent test-retest study is presented in Section 8.2. All scales score at or above 0.79 as measured by the standard correlation coefficient. All correlations are computed with a statistical significance of less than 0.001. Given the solid correlation between the scales across time, we

consider this to be solid evidence of temporal consistency. Our most recent test-retest study prior to the current study presented in this work is documented in [24].

16.3 Alternate Forms Reliability

As noted elsewhere in this work, alternate forms consistency is, in essence, a call for multiple versions of the same assessment for comparison across a population. TTI Success Insights views this approach to be cost prohibitive and view any potential insights gained through such a process to be limited in comparison to the capital outlay required. For this reason, we do not foresee, at this time, an alternate forms reliability study in our future.

16.4 Generalizability Theory

Our approach to Generalizability theory is outlined in Section 9. We note for the reader that the term generalizability is used in this context as a generalization of classical test theory versions of reliability, or internal consistency reliability to be more accurate. In other words, good scores based on this theory should not be interpreted to mean that the Style Insights assessment scores can be generalized across entire populations. That interpretation of generalizable should be considered as part of Relationships to External Variables validity evidence as discussed in Section 13 and summarized below in Section 16.7. Table 9.9 summarizes the results by presenting the Generalizability and Dependability coefficients. The scores range between 0.88 and 0.91 and may be interpreted in a similar fashion to internal consistency coefficients.

16.5 Test Content Validity Evidence

Section 2 covers the historical development from the early assessment development of Clarke and Cleaver through to Geier's version from the 1960s. The current version of the Style Insights assessment is essentially the same as Geier's final version. The historical studies are documented in this section.

16.6 Internal Structure Validity Evidence

Internal structure validity is approached in several ways in this manual. The first is the use of (corrected) item-total correlation as a proxy for item discrimination. This is reported in two ways. We use item-total correlation as our measure of item discrimination in Section 4 where we discuss item analysis, see [136]. We are essentially considering how discriminating the assessment items are at several levels of difficulty (the ability spectrum), in some sense a foreshadowing of the use of the slope of the item characteristic curves to do so on the continuum defined by the ability spectrum.

The next approach is a direct reporting of (corrected) item-total correlation coefficients for each item on each scale. We note that this does not directly establish a complete internal structure validity evidence argument, but it does lay a foundation. We also are using this as a diagnostic approach to identify items which may be underperforming. This information is presented in Figure 6.1 and Table 6.2. Based on the information presented in Table 6.1, borrowed from the APA, the vast majority of items score at or above the cutoff for moderate levels of information, with many items scoring above the large information cutoff.

It is noted that a handful of the items falls below the moderate level cutoff. Such items are noted and will be observed in the near term with an eye on replacement in the long term. The subject of replacement items for our assessments is discussed below in Section 16.10.

We have noted on several occasions in this manuscript that several of the tools are not directly applicable to our assessment data due to its forced-rank response format. We have argued that while the methods may not be completely directly applicable to our data, they may be applied to the individual scales, noting that the interpretation of the results is difficult. With this caveat noted, we use exploratory factor analysis (EFA), item response theory (IRT), and multivariable item response theory (MIRT) to aid our diagnostic approaches to assessment development and improvement.

We do this in two ways. First, we analyze our assessment data by extracting the individual scale responses from the forced-rank data. This is the part that the academics disapprove. We again note that we do so for diagnostic purposes with an eye on the future. The second approach is to utilize a data gathering system developed internal to TTI Success Insights allowing us to test individual items. We gather the individual item data on a likert-style scale whose data is useful in many of these statistical approaches.

Our approaches to EFA, IRT, and MIRT are presented in Sections 10, 11, and 12, respectively. EFA and IRT are applied to both the forced-rank data for diagnostic purposes and the likert-style data for testing and improvement purposes. The MIRT is currently in the form of a Thurstonian IRT approach as outlined in [30]. This implementation is in its early stages, but preliminary results are promising, albeit on small subsets of the overall assessment.

16.7 Relationships to External Variables Validity Evidence

Section 13 and Appendix B are dedicated to outlining current efforts in this area. For a brief synopsis, we are currently collecting demographic data on our U.S. based English assessment. Included in the subset of variables we are collecting is the O*NET job classification data. We are using this data and our Style Insights variables, and derivatives thereof, to model set inclusion. The corresponding probability model is analyzed for accuracy via the area under the receiver operating characteristic (ROC) curve. An area under the curve (AUC) score of greater than 0.50 corresponds to the model exceeding the usual base model generated by random selection.

It is our opinion that this approach provides superior information when compared to the usual approach of computing a simple validity coefficient, which really should be read as correlation coefficient. Given that correlation does not establish any form of causation, it would seem that the industries who use assessments in selection, retention, and hiring would ask for more than a correlation coefficient. A more complete argument for this case is presented in Section 3.3.

When the data were pulled for this manuscript, the job classification variable showed 220 cases with 100 or more respondents. This modeling approach has provided probability models with AUC scores greater than 0.50, most greater than 0.60, for 178 of the aforementioned 220 cases.

16.8 Response Processing Validity Evidence

A complete treatment of one of our major approaches to response processing validity evidence gathering is presented in Section 14. This is an area of research that the team at TTI Success Insights is particularly excited to resume as soon as possible. The major first step in our process is to use EEG to capture brain activity during the time a subject is responding to our assessment items or frames. We then combine the EEG output (as images of the brain) based on gamma wave activity and compare with statistical analysis taken from the general TTI Success Insights population on the same assessment items or frames.

The output is then analyzed by the neurology team and then discussed with the mathematical/analytic team where we compare notes. The protocols for this process have been published in a paper in late 2018, see [26]. In this paper we describe several potential response patterns that emerge and then we attempt to tie these results into the ideas of item difficulty and discrimination. Some examples of this joint analysis come from items that the study population generally shows agreement. In this setting, agreement is interpreted to mean that the brain images show agreement with the study participants' responses to an item or frame. In this case we expect to see relatively solid discrimination parameters in an IRT model, or reasonable levels of item-total correlation in the non-IRT cases, for a given item.

In other cases we see items that are intended to be reverse scored or keyed in which brain imaging shows confusion while mathematically we see low discrimination scores and potentially oddly distributed population response distributions. Overall, the idea is to use the mathematical analysis to identify the good (and not so good), while using the EEG generated brain images to identify any causes for potentially under-performing items. The combination is proving to be a quite solid approach to identifying information about our assessments and the individuals responding to our assessments.

As a final note, TTI Success Insights is quite proud of our merging of neurology and mathematical analysis for at least two reasons. The first is that we are unaware of any other assessment company in the market who are conducting neurological based studies. We believe that the analysis being conducted at TTI Success Insights is truly an industry leader at this time. The second reason is that according to the APA,

> 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. (pp. 76, [156])

While it is not completely certain we are witnessing the cognitive process directly, we are gathering information on how the individual's brain activity is changing before, during, and after seeing and responding to an assessment item.

16.9 Consequences of Testing Validity Evidence

The TTI Success Insights approach to consequences of testing validity evidence is outlined in Section 15. As is noted in this section, the primary responsibility for ensuring unintended consequences are evaluated lies with the test user. In the vast majority of the administrations of our assessments, TTI Success Insights is not the test user. Rather, we play the role of test developer and test administrator. This is not to say that we do not take into account the possible unintended consequences that may result through misuse or misinterpretation of our assessment data.

As summarized in Section 15, we focus on training those individual who are in the role of test user to ensure a high level of understanding of the appropriate uses and interpretations of the TTI Success Insights suite of assessments. This includes ongoing education for the network of value-added associates that make up the larger family of TTI Success Insights assessment distributors. Historically we have offered in-person courses on many aspects of our assessments from early stage "boot camp" to more advanced courses that tie together multiple assessments.

TTI Success Insights has been transitioning many education opportunities and resources to online platforms. As a result of the COVID-19 pandemic, all educational opportunities currently offered are online only.

16.10 TTI Success Insights Continual Improvement Process

Continual improvement is a key theme throughout this work as well as being an integral part of our day to day operations at TTI Success Insights. The entire team at TTI Success Insights has made a concerted effort over the years to be at the forefront of assessment industry quality standards and to provide assessments with sound research based foundations.

In the next, and final, section of the main body of this work, we outline some of the future projects we have outlined for the Style Insights assessment. On top of these forthcoming projects, we have a host of ongoing, periodic analyses and studies related to various aspects of the concepts of evidence of reliability and validity. Many have been noted elsewhere in this work and we do not wish to become overly repetitive. We only wish to highlight the continued commitment of TTI Success Insights' leadership and research teams to ensuring a continual improvement process is in-place and operating to ensure the highest standards are aimed for and met.

16.11 Overall Summary and Future Work

The TTI Success Insights research team has two main subgroups that work in close alignment, yet are quite different in their individual foci. One of the authors of this manuscript came to TTI Success Insights in 2011 with a goal of tying neurological research into the various aspects of validity. The other author came to TTI Success Insights in 2017 and has since focused on the more analytical aspects and psychometric properties of our assessments. One of our main goals has been, and will continue to be, to form a cohesive unit that merges these two research components into a single, powerful set of tools to aid our understanding of all elements of the TTI Success Insights suite of assessments.

One prime example of the potential fruits of this labor may be seen in the publication of a paper outlining the protocols for combining the neurological insights gained via EEG imaging and the mathematical insights coming from IRT and EFA models, see [26]. This paper, along a forthcoming follow-up paper detailing the results of an initial study combining the two disciplines, provides a snapshot of what is possible in this space. At the time of writing this manuscript, the authors are aware of no other psychometric assessment company that is going to this level of effort to uncover the relationships between the brain and the mathematics of these type assessments.

Our continual improvement process has been focused on three main concepts over the past several years. The first is the neurological, followed by the mathematical, and finally the merging of the two. There is no reason to suspect this to change. In fact, it is our intention to expand on all the areas discussed in this work and to continue to monitor the current trends in academia for advances or breakthroughs that may strengthen our understanding. We provide highlights of what this vision may provide for the future at TTI Success Insights.

As referenced above, continued research into the neurological aspects of our assessments are a major focus of our research team. We had multiple neurological studies planned and had just begun data collection on one when the world was interrupted by the COVID-19 pandemic in Spring, 2020. While there were no legal issues with our continued efforts to gather data of which we are aware, an abundance of caution forced us to suspend such studies due to the fact that the subject must be physically present and connected to the EEG equipment for us to gather the relevant data. As soon as it is possible for us to safely begin collecting EEG data again, it is our intention to resume our neurological studies. Our currently suspended study was intent on gathering data around the Motivation Insights™ assessment with an additional study scheduled for later in 2020 to do

essentially the same study on the Style Insights™ assessment. We are quite excited to get back into the lab, get these studies kicked-off, and see what knowledge we may gain.

There are three main areas of focus for the mathematical side of our research team: Exploratory Factor Analysis (EFA), Item Response Theory (IRT), and multivariable IRT. It has been noted in earlier in this manuscript that EFA and IRT are not directly applicable to the assessment format at hand. The forced-rank format generates a covariance matrix structure that, by definition, is singular. A singular matrix is not invertible, hence any attempt to model a forced-rank format assessment is useless. It is possible to separate the constructs from each other in the data, thereby removing the aforementioned singularity. However, as noted in [94], while this is mathematically possible, the interpretation of any results are questionable. Removing the physical connection of the constructs in the data does not remove their impact on each other. Hence, interpretation is difficult at best.

We have noted earlier that we do, in fact, employ these tactics in our study of our Style Insights assessment and its respondent data. The reader may be wondering why state this after just arguing that interpretation is difficult at best. The first part of the answer to this very relevant question is that while interpretation is difficult, it is not impossible and much valuable information may be gained to advance our understanding of our assessments. The second part of the answer is that we also continually test our assessment items, both currently in use items and potential replacements, in a likert-style format. In that case, the EFA and IRT models are completely appropriate. It is through this combination of approaches that we anticipate gaining the most information about how our assessments are performing.

In addition to the information gain, we also are moving towards advancing our analytic methodologies for all our assessments from the mostly archaic classical test theory approach to the more modern item response theory approach. In this way, we are attempting to modernize both our understanding of our assessment reliability, in the form of item and test information functions, and internal structure validity, via the item characteristic curves and their associated item discrimination parameters. Clearly this only holds in the likert-style format. Once again, we argue that solid information may be obtained in this format and employed in the forced-rank format to strengthen the forced-rank based assessments.

We also intend to advance our approach to the forced-rank format assessments in our suite by employing multivariable IRT models, such as the Thurtonian approach as outlined in [30], among many others. According to this approach, one may use a combination of Thurstone's Law of Comparative Judgment, confirmatory factor analysis, and item response theory to generate a multivariable IRT model for the forced-rank assessment data. While to date we have not generated a full Thurstonian IRT model for our Style Insights assessment, early results based on small subsets of the assessment are promising. As pointed out in [94], one of the criticisms of forced-rank data is that they are not applicable for interpersonal comparisons. We plan to alleviate this criticism by employing a full multivariable IRT model in the future.

One very important area of interest deserves special mention. That area is invariance in general, but specifically invariance of internal structure. In some settings this may be thought of as approaching the problem of disparate impact (Civil Rights Act of 1964, Title VII), also known more informally as adverse impact in the business community. TTI Success Insights has historically measured adverse impact on a periodic basis for most of our assessments. The current version is hosted at https://images.ttisi.com/wp-content/uploads/2019/07/02120329/TTISI_AdverseImpactStudy_2019.pdf. Archived versions of previous studies may be requested from the

TTI Success Insights research team. There is a scheduled updating of the differential effects associated with several of our assessments, due to be released Summer, 2021. As with many other areas of analysis, it is our intent to move studies of assessment invariance to a more modern analytical approach over the next few years. For example, we are in the early stages of assessing differential item functioning of our likert-style format assessments using graded response models.

It is a tall order to lay out all possible paths that our research may take over the next several years. The previous paragraphs outline a handful of the directions we plan to explore. Additionally, we plan to continue our exploration of regression modeling of our assessment variables against any external data we may be able to obtain, including expanding our use of the demographic data discussed in this manuscript. We also are discussing the possibility of an extended look at our overall population and potentially relevant subsets of our population to determine if our variables may are indicating any significant changes based on pre and post COVID-19 pandemic data. An early look at our Style Insights variables comparing a similar time period in 2019 and 2020 did not indicate such changes. However, hindsight shows us that our study may have been premature given the now yearlong (and growing) series of lockdowns, stay-at-home orders, job losses, and more.

A final short discussion around an area we have long been interested shall close out this summary. Given the vast geographic area that is the United States, many questions have come up over the years around ideas such as whether individuals in northeastern cities (New York, Boston, Philadelphia, etc.) are really different than individuals from rural upper midwest (Iowa, Minnesota, or the Dakotas) or west coast cities (Los Angeles, San Francisco, Seattle). Given that individuals living in New York City are nearly as close to London as they are to Los Angeles (a quick Google search shows nearly identical average nonstop flight times), it may be necessary to consider whether language usage differences in the very different geographic locations causes any issues in assessment interpretation.

We have outlined many ways in which the TTI Success Insights team is working to continue to produce the highest quality assessments available in the market. We employ cutting-edge, neurological research studies. We apply the latest theoretically based techniques in the mathematical and statistical analysis of our assessments. Finally, we have a team that is working diligently to merge the two disciplines to form a difficult to beat combined understanding of our assessments and the interpretations of them. The future is bright at TTI Success Insights.

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, Third Edition*. Wiley, New York, 2012.

[3] C. Angus. How accurate are personality tests. Scientific American, 10 2018.

[4] H.L. Ansbacher. Prescott Lecky's concept of resistance and his personality. *Journal of Clinical Psychology*, 37:791–794, 1965.

[5] 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.

[6] 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.

[7] Frank B. Baker. *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, New York, 1992.

[8] 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.

[9] 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.

[10] Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. The Guilford Press, New York, 2018.

[11] N. Barger and L. Kirby. *Introduction to Type and Change*. CPP, Inc., Mountain View, CA, 2004.

[12] 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.

[13] 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.

[14] R.D. Bock. A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16(4):21–33, 1997.

[15] Douglas G. Bonett and Thomas A. Wright. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, 2014.

[16] B. Bonnstetter. Farmer buying style study based on observable farmstead features. TTI Success Insights, 1979.

[17] Bill J. Bonnstetter, Ronald J. Bonnstetter, Dustin Hebets, and Tom F. Collura. Validation process for ipsative assessments, 2016. US Patent 9,060,702.

[18] Bill J. Bonnstetter, Ronald J. Bonnstetter, Dustin Hebets, and Tom F. Collura. Validation process for ipsative assessments, 2016. Canadian Patent No. 2,808,691.

[19] 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.

[20] R.J. Bonnstetter. Natural Versus Adapted DISC Graphs. White Paper, TTI Success Insights, 12 2020.

[21] 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.

[22] 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.

[23] Ronald J. Bonnstetter, Thomas F. Collura, and Dustin Hebets. Uncovering the Belief Behind the Action. *NeuroConnections*, 2012.

[24] Ronald J. Bonnstetter and Eric T. Gehrig. TTI Success Insights Style Insights 2016 Temporal Consistency Report. Technical report, TTI Success Insights, 4 2016.

[25] 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.

[26] Ronald J. Bonnstetter, Eric T. Gehrig, and Dustin Hebets. Response Process Validation Protocol using Neurophenomenological Gamma Asymmetry. *NeuroRegulation*, 2018.

[27] 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.

[28] Robert L. Brennan. *Generalizability Theory*. Statistics for Social Science and Public Policy. Springer, New York, 2001.

[29] Anna Brown. Doing less but getting more: Improving forced-choice measures with item response theory. *Assessment and Development Matters*, 2(1):21–25, 2010.

[30] Anna Brown and Albert Maydeu-Olivares. Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurements*, 71(3):460–502, 2011.

[31] Anna Brown and Albert Maydeu-Olivares. Fitting a Thurstonian IRT model to forced-choice data. *Behavioral Research Methods*, 44:1135–1147, 2012.

[32] Anna Brown and Albert Maydeu-Olivares. How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18:36–52, 2013.

[33] Anna Brown and Alberto Maydeu-Olivares. Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4):516–529, 2018.

[34] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910.

[35] R.H. Burros. The estimation of the discriminal dispersion in the method of successive intervals. *Psychometrika*, 20:299–305, 1955.

[36] J.M. Cattel. Mental tests and measurement. *Mind*, 15:373–380, 1890.

[37] W.F. Chaplin and K.E. Buckner. Self-ratings of personality: A naturalistic comparison of normative, ipsative, and idiothetic standards. *Journal of Personality*, 56:509–530, 1988.

[38] Eunseong Cho and Seonghoon Kim. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2):207–230, April 2015.

[39] Lee Anna Clark and David Watson. Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3):309–319, 1995.

[40] W.V. Clarke. The construction of an industrial selection personality test. *Journal of Psychology*, 1:379–394, 1956.

[41] W.V. Clarke. The personality profiles of life insurance agents. *Journal of Psychology*, 42:295–302, 1956.

[42] W.V. Clarke. The personality profiles of loan office managers. *Journal of Psychology*, 41:405–412, 1956.

[43] W.V. Clarke. The personality profiles of self-made company presidents. *Journal of Psychology*, 41:413–418, 1956.

[44] W.V. Clarke. The personality profiles of self-made company presidents. *Journal of Psychology*, 41:413–418, 1956.

[45] J. Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7:249–253, 1983.

[46] T. F. Collura, C. Zalaquett, and R. J. Bonnstetter. Seeing inside the client's mind. *Counseling Today*, 57(6):24–27, 2014.

[47] 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.

[48] T.F. Collura, R.J. Bonnstetter, D. Hebets, and B.J. Bonnstetter. Uncovering the Belief Behind the Action. *NeuroConnections*, Winter:20–23, 2012.

[49] T.F. Collura, R.J. Bonnstetter, and C.P. Zalaquett. Neurocounseling: Bridging Brain and Behavior. *Counseling Today*, 57(6):24–27, 2014.

[50] 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.

[51] 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.

[52] 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.

[53] International Test Commision. The ITC Guidelines for Translating and Adapting Tests (Second Edition). https://www.InTestCom.org, 2010.

[54] D.J. Cooke and C. Michie. An item response theory analysis of the hare psychopathy checklist-revised. *Psychological Assessment*, 9:3–14, 1997.

[55] Jose M. Cortina. What is Coefficient α? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78(1):98–104, 1993.

[56] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

[57] 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.

[58] 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.

[59] Robert F. DeVellis. *Scale Development: Theory and Applications.* Sage, Thousand Oaks, CA, 4th edition, 2017.

[60] William R. Dillon and Matthew Goldstein. *Multivariate Analysis: Methods and Applications.* Wiley Series in Probability and Mathematical Statistics, 1984.

[61] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412, 2014.

[62] 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.

[63] R. Eveleth. The myers-briggs personality test is pretty much worthless. Smithsonian Magazine, 3 2013.

[64] Carl F. Falk and Victoria Svalei. The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment*, 93(5):445–453, 2011.

[65] 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.

[66] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[67] Ronald A. Fisher. *Theory of Statistical Estimation.* John Wiley & Sons, New York, 1950. Proceedings of the Cambridge Philosophical Society 22 (1925) 700-725.

[68] 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.

[69] 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.

[70] Eric T. Gehrig. Classification of Serial Entrepreneurs via Logistic Regression: A Case Study. White Paper, October 2017.

[71] J.G. Geier and D.E. Downey. *Energetics of Personality: Defining a Self*. Aristos Publishing House/Geier Learning Systems, United States, 1989.

[72] J.G. Geier and D.E. Downey. *Personality Analysis*. Aristos Publishing House, United States, 1989.

[73] 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.

[74] 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.

[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 2: Testing and Assessment in Clinical and Counseling 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 3: Testing and Assessment in School Psychology and Education*, 750 First Street NE, Washington, D.C. 20002-4242, 2013. American Psychological Association.

[77] Richard L. Gorsuch. *Factor Analysis*. Lawrence Erlbaum Associates, 2nd edition, 1983.

[78] A. Grant. Goodbye to MBTI, the Fad That Won't Die. Psychology Today, 9 2013.

[79] Bert F. Green. In defense of measurement. *American Psychologist*, 33:664–670, 1978.

[80] James W. Grice. Computing and evaluating factor scores. *Psychological Methods*, 6(4):430–450, 2001.

[81] H. Gulliksen. *Theory of Mental Test Scores*. Wiley, New York, 1950.

[82] Louis Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 42(4):252–282, 1945.

[83] 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.

[84] Harry H. Harman. *Modern Factor Analysis*. University of Chicago Press, Chicago and London, third revised edition edition, 1976.

[85] D. Harnisch. 2015 style insights reliability study. Education and Human Resources, University of Nebraska, TTI Success Insights, 2015.

[86] 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.

[87] R. Hogan. *Personality and the Fate of Organizations*. Psychology Press, New York, 2015.

[88] W.H. Holtzman and S.B. Selss. The prediction of flying success by clinical analysis of test protocols. *Journal of Abnormal Social Psychology*, 49:485–490, 1954.

[89] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, Inc., 2000.

[90] C. Hoyt. Test reliability estimated by analysis of variance. *Psychometrika*, 6:153–160, 1941.

[91] TTI Success Insights. Disparate impact study. Unpublished white paper., 2019.

[92] R. W. B. Jackson and G. A. Ferguson. Studies on the reliability of tests. Technical report, University of Toronto, 1941.

[93] 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.

[94] 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.

[95] J.T.Lamiell. Individuals and the differences between them. In R. Hogan, J. Johnson, and S. Briggs, editors, *Handbook of Personality Psychology*. Academic Press, San Diego, 1997.

[96] C.G. Jung. *Psychological Types*. 1923. Translation by H. Godwyn Baynes.

[97] Robert M. Kaplan and Dennis P. Saccuzzo. *Psychological Testing: Principles, Applications, and Issues*. Wadsworth, Belmont, CA, 2001.

[98] E.L. Kelly and D.W. Fiske. *The Prediction of Performance in Clinical Psychology*. University of Michigan Press, Ann Arbor, MI, 1951.

[99] 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.

[100] 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.

[101] G. F. Kuder and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937.

[102] J.T. Lamiell. On the utility of looking in the "wrong" direction. *Journal of Personality*, 48:82–88, 1980.

[103] J.T. Lamiell. *The Psychology of Personality: An Epistemological Inquiry*. Columbia University Press, New York, 1987.

[104] 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.

[105] 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.

[106] P. Lecky. Review of Die psychische reeingflussing des sauglings. *Internationale Zeitschrift für Indivdualpsychologie*, 6:506–507, 1928.

[107] P. Lecky. The theory of self-consistency in personal problems. Paper presented at the meeting of the American College Personnel Association, 1935.

[108] P. Lecky. Preventing failure by removing resistance. Paper presented at the meeting of the New York Society for the Experimental Study of Education, 1938.

[109] P. Lecky. *Self-consistency*. The Shoe String Press, New York, 1945.

[110] P. Lecky. *Self-consistency: A Theory of Personality*. Island Press, New York, 2^{nd} edition, 1951. Edited and interpreted by F.C. Thorne.

[111] P. Lecky. *Self-consistency: A Theory of Personality*. Island Press Cooperation, Inc., New York, 1961. Edited and interpreted by F.C. Thorne, reprinted by Island Press Cooperation, Inc.

[112] Jane Loevinger. The attenuation paradox in test theory. *Psychological Bulletin*, 51:493–504, 1954.

[113] 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.

[114] F. M. Lord. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13:517–548, 1952.

[115] F. M. Lord. A theory of test scores. *Psychometric Monograph, (No. 7)*, 1952.

[116] Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale, NJ, 1980.

[117] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Reading, MA, 1968.

[118] 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.

[119] George A. Marcoulides. Generalizability theory. In Howard E. A. Tinsley and Steven D. Brown, editors, *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, chapter 18, pages 527–551. Academic Press, 2000.

[120] William M. Marston. *Emotions of Normal People*. Kegan Paul, Trench, Trubner, & Co. Ltd., New York, 1928.

[121] Albert Maydeu-Olivares. Limited information estimation and testing of thurstonian models for paried comparison data under multiple juedgment sampling. *Psychometrika*, 66(2):209–228, 2001.

[122] Albert Maydeu-Olivares and Ulf Böckeenholt. Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, 10(3):285–304, 2009.

[123] Albert Maydeu-Olivares and Anna Brown. Ordinal factor analysis of graded preference questionnaire data. *Multivariate Behavioral Research*, 45:935–974, 2010.

[124] Albert Maydeu-Olivares and Anna Brown. Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, 44:1135–1174, 2012.

[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] P.F. Merenda and W.V. Clarke. Self-description and personality measurement. *Journal of Clinical Psychology*, 21(1):52–56, 1965.

[128] Samuel Messik. Meaning and values in test evaluation: The science of ethics and assessment. *Educational Researcher*, 18(2):5–1, 1989.

[129] Samuel Messik. Validity of psychological assessment: Validation inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749, 1995.

[130] Michael L. Nering and Remo Ostini, editors. *Handbook of Polytomous Item Response Theory Models*. Routledge: Taylor & Francis Group, New York and London, 2010.

[131] Jum C. Nunnally. *Psychometric Theory*. McGraw-Hill, New York, 2nd edition, 1978.

[132] 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.

[133] Remo Ostini and Michael L. Nering. Polytomous Item Response Theory Models. In *Quantitative Applications in the Social Sciences*, volume 144. Sage Publications, 2006.

[134] 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.

[135] 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.

[136] 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.

[137] Gjalt-Jorn Y. Peters. The alpha and omega of scale reliability and validity. *The European Health Psychologist*, 16(2):56–69, April 2016.

[138] 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.

[139] G. Rasch. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: The Danish Institute for Educational Research, 1960. G. Leunbach, Trans.

[140] 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.

[141] 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.

[142] Tenko Raykov and George A. Marcoulides. *Introduction to Psychometric Theory*. Routledge, New York, 2011.

[143] William Revelle and Richard E. Zinbarg. Coefficients alpha, beta, omega, and the glb: comments on sijtsma. *Psychometrika*, 74(1):145–154, 2009.

[144] 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.

[145] Malcom Ritter. Russian mathematician rejects $1 million prize. https://phys.org/news/2010-07-russian-mathematician-million-prize.html, 2010. Phys.org.

[146] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17, 1969.

[147] F. Samejima. A general model for free response data. *Psychometrika Monograph Supplement*, 18, 1972.

[148] 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).

[149] 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.

[150] F. Samejima. Evaluation of the mathematical models for ordered polychotomous responses. *Behvariometrika*, 23:17–35, 1996.

[151] F. Samejima. Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, 62:471–493, 1997.

[152] Richard J. Shavelson and Noreen M. Webb. *Generalizability Theory: A Primer*. Sage, Thousand Oaks, CA, 1991.

[153] 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.

[154] Lorrie A. Shepard. Evaluating test validity. In L. Darling-Hammon, editor, *Review of Research in Education*, volume 19, pages 405–450. AERA, Washington, D.C., 1993.

[155] 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.

[156] 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.

[157] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[158] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910.

[159] 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.

[160] M.J. Stevens. Prescott Lecky: Pioneer in consistency theory and cognitive therapy. *Journal of Clinical Psychology*, 48(6):807–811, 1992.

[161] Yoshio Takane and Jan de Leeuw. On the Relationship Between Item Response Theory and Factor Analysis of Discretized Variables. *Psychometrika*, 52:393–408, 1987.

[162] 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.

[163] Inc. Target Training International. Employee success prediction system, 1996. US Patent 5,551,880.

[164] Inc. Target Training International. Network based document distribution method (ids), 2007. US Patent 7,249,372.

[165] Inc. Target Training International. Position analysis system and method (benchmark), 2007. US Patent 7,184,969.

[166] Ltd. Target Training International. 2017 talent insights reliability study. Unpublished research study, 2017.

[167] F.C. Thorne. A biographical sketch. In P. Lecky, editor, *Self-consistency: A Theory of Personality*, pages 9–19. Island Press, New York, 1951.

[168] L. L. Thurstone. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16:433–451, 1925.

[169] L. L. Thurstone. *Multiple Factor Analysis*. University of Chicago Press, Chicago, 1947.

[170] L. R. Tucker. Maximum validity of a test with equivalent items. *Psychometrika*, 11:1–13, 1952.

[171] S.R. Wallace, W.V. Clarke, and R.J. Dry. The Activity Vector Analysis as a selector of life insurance salesmen. *Personnel Psychology*, 49:37–345, 1956.

[172] D. Warburton. *The importance of finding the real person at work*. Department of Psychology, University of Reading, Whiteknights, Reading, UK, 1998.

[173] D.M. Warburton and J.I. Suiter. Discovering the person at work. In B. Bonnstetter, J. Suiter, and R. Widrick, editors, *The Universal Language DISC: A Reference Manual*. Target Training International, Inc., Scottsdale, AZ, 1993.

[174] D.M. Warburton and J.I. Suiter. Discovering the person at work. In B. Bonnstetter, J. Suiter, and R. Widrick, editors, *The Universal Language DISC: A Reference Manual*. Target Training International, Inc., Scottsdale, AZ, 1999. Reprint and update of the 1993 version.

[175] 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.

[176] 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.

[177] 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.

[178] 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.

A 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 [132]. 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 [53], 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.

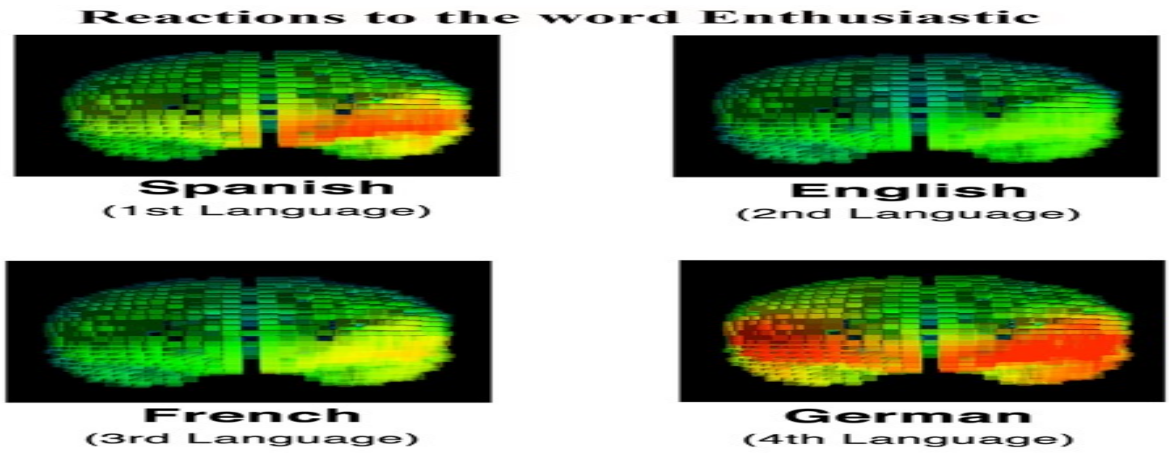

Figure A.1: Emotional response to *enthusiastic*

Before going on to the guidelines, we present a short motivation for the importance of having validated test adaptation protocols. Figure A.1 shows the emotional response of a multi-lingual

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 [52].

> 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.

A.1 Pre-conditions

There are three pre-condition guidelines in [53], 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 [53], 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.

A.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.

A.3 Confirmation Guidelines

While there are only four main guidelines in the Confirmation section of [53], 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.

[53] 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 [53] 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 [53],

> 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.

[53] 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.

[53] 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**, [53] 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 [155]. It is noted in [53] 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 [53]:

> 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, [53] suggests essentially following the standards set forth in [5] as well as in [156]. 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.

[53] 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, [53] 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?

A.4 Administration Guidelines

While there are only two main guidelines in the Administration section of [53], 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.**

A.5 Score Scales and Interpretation Guidelines

While there are only two main guidelines in the Score Scales and Interpretation section of [53], 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 [53], 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 [53] 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.

A.6 Documentation Guidelines

Finally, there are only two main guidelines in the Documentation section of [53]. 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.

B Relationships to External Variables Results

This Appendix is reserved for presenting the current results from our logistic regression approach to occupational title identification through the Style Insights variables. The general idea behind the approach to the current section is outlined in Section 13. As a very short refresher, the approach taken is to use a set membership/identification approach that relies on application of the logistic regression modeling procedure. This procedure is used to identify the probability of membership in the target set (i.e., the job classification) when the target set is included in a random sample of the general population. In this setting, the term general population refers to the larger TTI Success Insights overall database.

B.1 Demographic Breakdown of the Data

The TTI Success Insights demographic collection includes several generic variables that may be of interest in the setting of measuring disparate impact, e.g., gender, disability or veteran status, ethnicity, etc., as well as the use of the O*NET job classification nomenclature. Certain variables are collected on all individuals (e.g., gender for purposes of identifying pronoun usage for report generation, see Remark B.1 below). Other variables are collected solely on a voluntary basis (e.g., disability status or ethnicity, see Note B.2 below). Finally, some variables are not presented to certain internal assessment classifications (e.g., an assessment generated for use as part of a hiring process, see Note B.3 below). Finally, it should be noted that any and all demographic information is collected for use for research purposes only, the lone exception being noted in Remark B.1. No persons or entities outside TTI Success Insights have access to any demographic information collected during this process.

Remark B.1. *Currently, information gathering questions precede all questionnaires collected via the TTI Success Insights Internet Delivery System (IDS). The nature of these questions is to gather such information as name, email, and gender, to name a few. Some of this information is used to generate unique identifiers during the data storage process. Others are used for report generation, i.e., selecting the appropriate gendered pronoun for use in the report. At the time of writing of this manual, TTI Success Insights was in the process of generating an up-to-date process for providing alternate pronoun options for individuals falling outside the male-female binary gender choice. A more complete description of this process is scheduled to be incorporated in this manual upon completed incorporation.*

Remark B.2. *All demographic variables currently collected in the TTI Success Insights assessment process are done so on a voluntary basis. Any assessment respondent may refuse to answer any or all of the demographic questions presented. The lone exception is noted in Remark B.1. Technically speaking, gender is not collected as part of the demographics collection. It has been historically collected as part of the information used to generate reports for the individual.*

Remark B.3. *Given that many of the TTI Success Insights assessment users participate in certain subsets of hiring processes with their clients, an option is provided in the TTI Success Insights system to select a "hiring link" when generating a series of assessments for a given client. When this hiring link is selected, certain demographic information is no longer offered as part of the collection process. For example, age and ethnicity are not longer offered as part of the demographic collection. Further, as noted above, no information collected by TTI Success Insights is available to any individual or entity outside TTI Success Insights.*

A brief description of the database and subsets used in this study are provided in the following tables. The data analyzed were collected between April 01, 2018 and August 31, 2020. The data set contains 549,782 total respondents.

Table B.1: TTI Success Insights Demographics
Gender, N = 549,782

Variable	Sample Size	Percentage
Male	292,174	53.14%
Female	257,608	46.86%

Table B.2: TTI Success Insights Demographics
Ethnicity, N = 155,892

Variable	Sample Size	Percentage
Caucasian	117,253	75.21%
Hispanic or Latino	11,769	7.55%
African American	9,867	6.33%
Asian	7,247	4.65%
Two or More Races	4,743	3.04%
Other	3,865	2.48%
American Indian or Alaskan Native	681	0.44%
Pacific Islander	467	0.30%

Table B.3: TTI Success Insights Demographics
Decade Born, N = 125,750

Variable	Sample Size	Percentage
1930-1940	80	0.06%
1940-1950	918	0.73%
1950-1960	8,511	6.77%
1960-1970	24,670	19.62%
1970-1980	30,732	24.44%
1980-1990	34,591	27.51%
1990-2000	24,346	19.36%
2000-2010	1,902	1.51%

Table B.4: TTI Success Insights Demographics
Education (US), N = 155,758

Variable	Sample Size	Percentage
No schooling	115	0.07%
Nursery School to 8th grade	276	0.18%
9th, 10th, or 11th grade	1,093	0.70%
12th grade, no diploma	511	0.33%
High School (or higher)	11,769	7.56%
Some college credit, but less than 1 year	7,032	5.51%
1 or more years of college, no degree	19,262	12.37%
Associates Degree (e.g., AA or AS)	13,243	8.50%
Bachelor's Degree	66,389	42.62%
Master's Degree	28,186	18.10%
Professional degree (e.g. MD, DDS, DVM, LLB, JD)	5,041	3.24%
Doctorate degree (e.g. PhD, EdD)	2,841	1.82%

Table B.5: TTI Success Insights Demographics
Employment Status, N = 157,419

Variable	Sample Size	Percentage
Employed	125,371	79.64%
Self-employed	13,457	8.55%
Out looking	10,708	6.80%
Student	6,087	3.87%
Retired	786	0.50%
Homemaker	642	0.41%
Out, not looking	304	0.19%
Unable	64	0.04%

Table B.6: TTI Success Insights Demographics
Disability Status, N = 155,472

Variable	Sample Size	Percentage
Yes	2,433	1.56%
No	153,039	98.44%

Table B.7: TTI Success Insights Demographics
Veteran Status, N = 155,472

Variable*	Sample Size	Percentage
Not a Veteran	147,693	94.51%
Other Veteran	6,716	4.30%
Disabled Veteran	1,553	0.99%
Vietnam Veteran	311	0.20%

Note B.4. $*$ *There are more classes of Veteran Status than are presented here. A decision was made to collect the above four cases rather than the extended version. The interested reader may consult the U.S. Department of Veterans Affairs,* https://www.va.gov*, or the relevant sections of the United States Code (U.S.C.) pertaining to definitions of veteran status.*

B.2 Model Results: No Holdout Samples

Table B.8: Acute Care Nurses Sample
Area Under Curve, N = 136

Sample	Variable	AUC	Brier
1	Frequent Change	0.57	0.28

Table B.9: Advertising & Promotions Managers Sample
Area Under Curve, N = 220

Sample	Variable	AUC	Brier
1	Consistency	0.64	0.25

Table B.10: Agricultural Workers, All Other Sample
Area Under Curve, N = 320

Sample	Variable	AUC	Brier
1	People Oriented	0.58	0.27

Table B.11: Airline Pilots, Copilots, & Flight Engineers Sample
Area Under Curve, N = 212

Sample	Variable	AUC	Brier
1	Frequent Change, Consistency	0.64	0.24

Table B.12: Architectural & Engineering Managers Sample
Area Under Curve, N = 167

Sample	Variable	AUC	Brier
1	Ad C	0.68	0.22

Table B.13: Art Directors Sample
Area Under Curve, N = 201

Sample	Variable	AUC	Brier
1	People Oriented	0.62	0.27

Table B.14: Assemblers & Fabricators, All Other Sample
Area Under Curve, N = 411

Sample	Variable	AUC	Brier
1	FUFT*, Following Policy	0.64	0.26

Note B.5. $*$ *Follow Up and Follow Through*

Table B.15: Auditors Sample
Area Under Curve, N = 148

Sample	Variable	AUC	Brier
1	Organized Workplace, Competitiveness, Consistency	0.65	0.22

Table B.16: Automotive Body & Related Repairers Sample
Area Under Curve, N = 175

Sample	Variable	AUC	Brier
1	Ad S	0.58	0.27

Table B.17: Automotive Engineers Sample
Area Under Curve, N = 128

Sample	Variable	AUC	Brier
1	People Oriented, Nat C	0.69	0.21

Table B.18: Automotive Service Technicians & Mechanics Sample
Area Under Curve, N = 120

Sample	Variable	AUC	Brier
1	Consistency, FUFT*	0.64	0.26

Note B.6. ∗ *Follow Up and Follow Through*

Table B.19: Baristas Sample
Area Under Curve, N = 109

Sample	Variable	AUC	Brier
1	Frequent Change*, Nat I	0.62	0.24

Note B.7. ∗ *Statistical Significance at* 0.90

Table B.20: Bartenders Sample
Area Under Curve, N = 110

Sample	Variable	AUC	Brier
1	FIWO*	0.72	0.21

Note B.8. ∗ *Frequent Interaction With Others*

Table B.21: Bill & Account Collectors Sample
Area Under Curve, N = 161

Sample	Variable	AUC	Brier
1	Organized Workplace, Ad S	0.64	0.24

Table B.22: Billing & Posting Clerks Sample
Area Under Curve, N = 126

Sample	Variable	AUC	Brier
1	Competitiveness	0.69	0.23

Table B.23: Budget Analysts Sample
Area Under Curve, N = 125

Sample	Variable	AUC	Brier
1	Ad I	0.62	0.27

Table B.24: Building Cleaning Workers, All Other Sample
Area Under Curve, N = 112

Sample	Variable	AUC	Brier
1	Frequent Change	0.58	0.26

Table B.25: Business Intelligence Analysts Sample
Area Under Curve, N = 287

Sample	Variable	AUC	Brier
1	Ad C	0.57	0.27

Table B.26: Carpenters Sample
Area Under Curve, N = 368

Sample	Variable	AUC	Brier
1	Analysis of Data, Nat C	0.58	0.26

Table B.27: Cashiers Sample
Area Under Curve, N = 155

Sample	Variable	AUC	Brier
1	Consistency, Customer Relations	0.66	0.23

Table B.28: Chemists Sample
Area Under Curve, N = 126

Sample	Variable	AUC	Brier
1	FIWO*	0.63	0.26

Note B.9. $*$ *Frequent Interaction With Others*

Table B.29: Child, Family, & School Social Workers Sample
Area Under Curve, N = 247

Sample	Variable	AUC	Brier
1	Customer Relations, FUFT*	0.59	0.25

Note B.10. ∗ *Follow Up and Follow Through*

Table B.30: Chiropractors Sample
Area Under Curve, N = 317

Sample	Variable	AUC	Brier
1	Ad S, FUFT*	0.60	0.25

Note B.11. ∗ *Follow Up and Follow Through*

Table B.31: Clergy Sample
Area Under Curve, N = 317

Sample	Variable	AUC	Brier
1	FIWO*	0.63	0.26

Note B.12. ∗ *Frequent Interaction With Others*

Table B.32: Combined Food Preparation & Serving Workers, Including Fast Food Sample
Area Under Curve, N = 317

Sample	Variable	AUC	Brier
1	Urgency	0.63	0.25

Table B.33: Compliance Managers Sample
Area Under Curve, N = 215

Sample	Variable	AUC	Brier
1	Analysis of Data	0.62	0.27

Table B.34: Computer & Information Research Scientists Sample
Area Under Curve, N = 134

Sample	Variable	AUC	Brier
1	Urgency, Organized Workplace	0.66	0.24

Table B.35: Computer Network Support Specialists Sample
Area Under Curve, N = 306

Sample	Variable	AUC	Brier
1	Versatility	0.64	0.24

Table B.36: Computer Programmers Sample
Area Under Curve, N = 487

Sample	Variable	AUC	Brier
1	Ad I	0.64	0.25

Table B.37: Computer Systems Analysts Sample
Area Under Curve, N = 360

Sample	Variable	AUC	Brier
1	Ad C	0.66	0.23

Table B.38: Construction Laborers
Area Under Curve, N = 114

Sample	Variable	AUC	Brier
1	Ad C	0.62	0.25

Table B.39: Cooks, All Other
Area Under Curve, N = 121

Sample	Variable	AUC	Brier
1	Urgency	0.62	0.24

Table B.40: Counselors, All Other
Area Under Curve, N = 240

Sample	Variable	AUC	Brier
1	Consistency, Nat S	0.60	0.25

Table B.41: Counter & Rental Clerks
Area Under Curve, N = 127

Sample	Variable	AUC	Brier
1	Consistency, Ad S	0.68	0.22

Table B.42: Credit Analysts
Area Under Curve, N = 127

Sample	Variable	AUC	Brier
1	People Oriented	0.58	0.26

Table B.43: Demonstrators & Product Planners
Area Under Curve, N = 477

Sample	Variable	AUC	Brier
1	Following Policy	0.65	0.23

Table B.44: Dentists, All Other Specialists
Area Under Curve, N = 146

Sample	Variable	AUC	Brier
1	Customer Relations	0.61	0.24

Table B.45: Dentists, General
Area Under Curve, N = 386

Sample	Variable	AUC	Brier
1	Nat S	0.58	0.27

Table B.46: Designers, All Other
Area Under Curve, N = 433

Sample	Variable	AUC	Brier
1	Ad S	0.56	0.27

Table B.47: Directors, Religious Activities & Education
Area Under Curve, N = 199

Sample	Variable	AUC	Brier
1	Nat D	0.58	0.29

Table B.48: Door-to-Door Sales Workers, News & Street Vendors, & Related Workers
Area Under Curve, N = 199

Sample	Variable	AUC	Brier
1	Ad D, Ad C	0.70	0.23

Table B.49: Driver-Sales Workers
Area Under Curve, N = 169

Sample	Variable	AUC	Brier
1	Nat D, Urgency	0.62	0.24

Table B.50: Electrical Engineers
Area Under Curve, N = 169

Sample	Variable	AUC	Brier
1	Ad I	0.62	0.26

Table B.51: Elementary School Teachers, Except Special Education
Area Under Curve, N = 363

Sample	Variable	AUC	Brier
1	Ad S, People Oriented	0.69	0.22

Table B.52: Emergency Medical Technician & Paramedics
Area Under Curve, N = 139

Sample	Variable	AUC	Brier
1	Ad S	0.57	0.26

Table B.53: Engineers, All Other
Area Under Curve, N = 354

Sample	Variable	AUC	Brier
1	Ad C, People Oriented	0.64	0.23

Table B.54: Entertainers and Performers, Sports & Related Workers, All Other
Area Under Curve, N = 104

Sample	Variable	AUC	Brier
1	FUFT*	0.62	0.25

Note B.13. $*$ *Follow Up and Follow Through*

Table B.55: Environmental Engineers
Area Under Curve, N = 134

Sample	Variable	AUC	Brier
1	Frequent Change	0.60	0.25

Table B.56: Environmental Scientists & Specialists, Including Health
Area Under Curve, N = 110

Sample	Variable	AUC	Brier
1	Versatility	0.64	0.23

Table B.57: Family & General Practitioners
Area Under Curve, N = 176

Sample	Variable	AUC	Brier
1	Nat D	0.59	0.29

Table B.58: Financial Clerks, All Other
Area Under Curve, N = 110

Sample	Variable	AUC	Brier
1	Competitiveness	0.65	0.25

Table B.59: Financial Managers
Area Under Curve, N = 484

Sample	Variable	AUC	Brier
1	People Oriented	0.63	0.23

Table B.60: First-Line Supervisors of Construction Trades & Extraction Workers
Area Under Curve, N = 279

Sample	Variable	AUC	Brier
1	Ad D, Ad C	0.64	0.23

Table B.61: First-Line Supervisors of Non-Retail Sales Workers
Area Under Curve, N = 338

Sample	Variable	AUC	Brier
1	Urgency, Consistency	0.65	0.23

Table B.62: First-Line Supervisors of
Office & Administrative Support Workers
Area Under Curve, N = 191

Sample	Variable	AUC	Brier
1	Nat S	0.59	0.29

Table B.63: First-Line Supervisors of Police & Detectives
Area Under Curve, N = 130

Sample	Variable	AUC	Brier
1	FIWO*	0.65	0.23

Note B.14. ∗ *Frequent Interaction With Others*

Table B.64: First-Line Supervisors of Production & Operating Workers
Area Under Curve, N = 208

Sample	Variable	AUC	Brier
1	FUFT*	0.58	0.26

Note B.15. ∗ *Follow Up and Follow Through*

Table B.65: First-Line Supervisors of Transportation &
Material-Moving Machine & Vehicle Operators
Area Under Curve, N = 171

Sample	Variable	AUC	Brier
1	Ad C	0.59	0.29

Table B.66: Food Preparation & Serving Related Workers, All Other
Area Under Curve, N = 140

Sample	Variable	AUC	Brier
1	Ad D, Ad C	0.61	0.24

Table B.67: Food Processing Workers, All Other
Area Under Curve, N = 144

Sample	Variable	AUC	Brier
1	FIWO*	0.57	0.24

Note B.16. $*$ *Frequent Interaction With Others*

Table B.68: Fundraisers
Area Under Curve, N = 172

Sample	Variable	AUC	Brier
1	Nat I	0.62	0.29

Table B.69: Graphic Designers
Area Under Curve, N = 449

Sample	Variable	AUC	Brier
1	Ad D	0.61	0.28

Table B.70: Ground Maintenance Workers, All Other
Area Under Curve, N = 138

Sample	Variable	AUC	Brier
1	Versatility	0.62	0.18

Table B.71: Hairdressers, Hairstylists, & Cosmetologists
Area Under Curve, N = 115

Sample	Variable	AUC	Brier
1	CR, OW*	0.64	0.15

Note B.17. ∗ *Customer Relations, Organized Workplace*

Table B.72: Health Technologists & Technicians, All Other
Area Under Curve, N = 122

Sample	Variable	AUC	Brier
1	Customer Relations	0.59	0.19

Table B.73: Healthcare Practitioners & Technical Workers, All Other
Area Under Curve, N = 327

Sample	Variable	AUC	Brier
1	Ad S	0.55	0.33

Table B.74: Heating, Air Conditioning, & Refrigeration Mechanics & Installers
Area Under Curve, N = 107

Sample	Variable	AUC	Brier
1	FIWO	0.59	0.20

Note B.18. ∗ *Frequent Interaction With Others*

Table B.75: Industrial Production Managers
Area Under Curve, N = 196

Sample	Variable	AUC	Brier
1	Ad C, Nat D	0.63	0.21

Table B.76: Information Security Analysts
Area Under Curve, N = 153

Sample	Variable	AUC	Brier
1	FIWO*	0.58	0.25

Note B.19. ∗ *Frequent Interaction With Others*

Table B.77: Installation, Maintenance, & Repair Workers, All Other
Area Under Curve, N = 498

Sample	Variable	AUC	Brier
1	Frequent Change	0.55	0.27

Table B.78: Insurance Policy Processing Clerks
Area Under Curve, N = 105

Sample	Variable	AUC	Brier
1	Nat S	0.68	0.23

Table B.79: Insurance Underwriters
Area Under Curve, N = 477

Sample	Variable	AUC	Brier
1	Nat D	0.55	0.31

Table B.80: Interior Designers
Area Under Curve, N = 242

Sample	Variable	AUC	Brier
1	Customer Relations, Nat I	0.64	0.25

Table B.81: Investment Underwriters
Area Under Curve, N = 126

Sample	Variable	AUC	Brier
1	People Oriented	0.56	0.22

Table B.82: Legal Support Workers, All Other
Area Under Curve, N = 365

Sample	Variable	AUC	Brier
1	Ad D	0.56	0.31

Table B.83: Librarians
Area Under Curve, N = 123

Sample	Variable	AUC	Brier
1	Ad S	0.67	0.27

Table B.84: Licensed Practical & Licensed Vocational Nurses
Area Under Curve, N = 125

Sample	Variable	AUC	Brier
1	Ad S	0.66	0.27

Table B.85: Life Scientists, All Other
Area Under Curve, N = 114

Sample	Variable	AUC	Brier
1	Nat D	0.56	0.25

Table B.86: Life, Physical, & Social Science Technicians, All Other
Area Under Curve, N = 146

Sample	Variable	AUC	Brier
1	Nat D	0.55	0.32

Table B.87: Logistics Managers
Area Under Curve, N = 241

Sample	Variable	AUC	Brier
1	People Oriented	0.61	0.21

Table B.88: Maintenance & Repair Workers, General
Area Under Curve, N = 261

Sample	Variable	AUC	Brier
1	People Oriented	0.59	0.25

Table B.89: Manufacturing Engineers
Area Under Curve, N = 231

Sample	Variable	AUC	Brier
1	Organized Workplace	0.65	0.25

Table B.90: Medical Assistants
Area Under Curve, N = 150

Sample	Variable	AUC	Brier
1	Nat S	0.61	0.28

Table B.91: Meeting, Convention, & Event Planners
Area Under Curve, N = 125

Sample	Variable	AUC	Brier
1	People Oriented	0.63	0.25

Table B.92: Military Officer Special &
Tactical Operations Leaders, All Other
Area Under Curve, N = 130

Sample	Variable	AUC	Brier
1	People Oriented	0.62	0.21

Table B.93: Network & Computer Systems Administrators
Area Under Curve, N = 386

Sample	Variable	AUC	Brier
1	Consistency	0.63	0.26

Table B.94: Nurse Practitioners
Area Under Curve, N = 226

Sample	Variable	AUC	Brier
1	People Oriented	0.66	0.25

Table B.95: Office Clerks, General
Area Under Curve, N = 177

Sample	Variable	AUC	Brier
1	Nat D, Nat S	0.72	0.21

Table B.96: Optometrists
Area Under Curve, N = 296

Sample	Variable	AUC	Brier
1	FC, CR*	0.65	0.26

Note B.20. ∗ *Frequent Change, Customer Relations*

Table B.97: Payroll & Timekeeping Clerks
Area Under Curve, N = 130

Sample	Variable	AUC	Brier
1	FIWO, Comp*	0.70	0.20

Note B.21. ∗ *Frequent Interaction With Others, Competitiveness*

Table B.98: Personal Care and Service Workers, All Other
Area Under Curve, N = 130

Sample	Variable	AUC	Brier
1	Ad C, Nat S	0.65	0.21

Table B.99: Pharmacists
Area Under Curve, N = 287

Sample	Variable	AUC	Brier
1	PO*, Ad I	0.62	0.21

Note B.22. ∗ *People Oriented*

Table B.100: Physical Therapists
Area Under Curve, N = 363

Sample	Variable	AUC	Brier
1	FC, PO*	0.62	0.26

Note B.23. ∗ *Frequent Change, People Oriented*

Table B.101: Physician Assistants
Area Under Curve, N = 101

Sample	Variable	AUC	Brier
1	Urgency	0.66	0.31

Table B.102: Physicians & Surgeons, All Other
Area Under Curve, N = 412

Sample	Variable	AUC	Brier
1	Ad I	0.57	0.25

Table B.103: Police Patrol Officers
Area Under Curve, N = 106

Sample	Variable	AUC	Brier
1	Ad I	0.58	0.25

Table B.104: Preschool Teachers, Except Special Education
Area Under Curve, N = 133

Sample	Variable	AUC	Brier
1	PO*, Ad S	0.73	0.19

Note B.24. $*$ *People Oriented*

Table B.105: Probation Officers & Correctional Treatment Specialists
Area Under Curve, N = 108

Sample	Variable	AUC	Brier
1	Ad S	0.58	0.31

Table B.106: Procurement Clerks
Area Under Curve, N = 113

Sample	Variable	AUC	Brier
1	Nat S	0.64	0.27

Table B.107: Producers
Area Under Curve, N = 155

Sample	Variable	AUC	Brier
1	Ad I	0.58	0.31

Table B.108: Production, Planning, & Expediting Clerks
Area Under Curve, N = 101

Sample	Variable	AUC	Brier
1	Nat D*	0.56	0.32

Note B.25. $*$ *Model significant at* 90% *confidence level.*

Table B.109: Protective Service Workers
Area Under Curve, N = 132

Sample	Variable	AUC	Brier
1	Versatility	0.58	0.25

Table B.110: Public Relations Fundraising Managers
Area Under Curve, N = 205

Sample	Variable	AUC	Brier
1	Ad I	0.66	0.26

Table B.111: Public Relations Specialists
Area Under Curve, N = 264

Sample	Variable	AUC	Brier
1	AoD, Cons*	0.66	0.26

Note B.26. ∗ *Analysis of Data, Consistency*

Table B.112: Purchasing Agents, Except Wholesale, Retail, & Farm Products
Area Under Curve, N = 134

Sample	Variable	AUC	Brier
1	OW*, Nat S	0.60	0.25

Note B.27. ∗ *Organized Workplace*

Table B.113: Purchasing Managers
Area Under Curve, N = 198

Sample	Variable	AUC	Brier
1	Ad C	0.57	0.33

Table B.114: Quality Control Systems Managers
Area Under Curve, N = 278

Sample	Variable	AUC	Brier
1	PO*, Nat C	0.60	0.21

Note B.28. ∗ *People Oriented*

Table B.115: Real Estate Brokers
Area Under Curve, N = 471

Sample	Variable	AUC	Brier
1	Consistency, Ad C	0.66	0.20

Table B.116: Receptionists & Information Clerks
Area Under Curve, N = 213

Sample	Variable	AUC	Brier
1	Urg, PO*	0.73	0.19

Note B.29. ∗ *Urgency, People Oriented*

Table B.117: Regulatory Affairs Managers
Area Under Curve, N = 103

Sample	Variable	AUC	Brier
1	Nat C	0.59	0.26

Table B.118: Risk Management Specialists
Area Under Curve, N = 331

Sample	Variable	AUC	Brier
1	Ad I	0.55	0.25

Table B.119: Roofers
Area Under Curve, N = 314

Sample	Variable	AUC	Brier
1	Ad D	0.58	0.23

Table B.120: Secondary Teachers, Except Special & Career, Technical Education
Area Under Curve, N = 130

Sample	Variable	AUC	Brier
1	Ad S	0.60	0.31

Table B.121: Secretaries & Administrative Assistants, Except Legal, Medical, & Executive
Area Under Curve, N = 387

Sample	Variable	AUC	Brier
1	Nat D, Nat S	0.70	0.22

Table B.122: Securities, Commodities, & Financial Services Agents
Area Under Curve, N = 183

Sample	Variable	AUC	Brier
1	Frequent Change	0.68	0.20

Table B.123: Social & Human Service Assistants
Area Under Curve, N = 171

Sample	Variable	AUC	Brier
1	PO*, Ad S	0.64	0.24

Note B.30. ∗ *People Oriented*

Table B.124: Social Workers, All Other
Area Under Curve, N = 493

Sample	Variable	AUC	Brier
1	PO*, Ad S	0.64	0.25

Note B.31. ∗ *People Oriented*

Table B.125: Software Developers, Systems Software
Area Under Curve, N = 166

Sample	Variable	AUC	Brier
1	Nat I	0.61	0.23

Table B.126: Software Quality Assurance Engineers & Testers
Area Under Curve, N = 166

Sample	Variable	AUC	Brier
1	Nat C	0.61	0.23

Table B.127: Special Education Teachers, All Other
Area Under Curve, N = 115

Sample	Variable	AUC	Brier
1	Ad S	0.63	0.30

Table B.128: Supply Chain Managers
Area Under Curve, N = 439

Sample	Variable	AUC	Brier
1	Comp*, Ad C	0.62	0.22

Note B.32. ∗ *Competitiveness*

Table B.129: Surgeons
Area Under Curve, N = 149

Sample	Variable	AUC	Brier
1	CR*, Nat I	0.62	0.22

Note B.33. ∗ *Customer Relations*

Table B.130: Tax Preparers
Area Under Curve, N = 163

Sample	Variable	AUC	Brier
1	Versatility	0.61	0.24

Table B.131: Teacher Assistants
Area Under Curve, N = 116

Sample	Variable	AUC	Brier
1	Urgency	0.70	0.28

Table B.132: Therapists, All Other
Area Under Curve, N = 251

Sample	Variable	AUC	Brier
1	Ad S	0.59	0.31

Table B.133: Training & Development Managers
Area Under Curve, N = 290

Sample	Variable	AUC	Brier
1	Organized Workplace	0.60	0.31

Table B.134: Training & Development Specialists
Area Under Curve, N = 426

Sample	Variable	AUC	Brier
1	Organized Workplace	0.60	0.31

Table B.135: Transportation Managers
Area Under Curve, N = 113

Sample	Variable	AUC	Brier
1	People Oriented	0.57	0.22

Table B.136: Waiters & Waitresses
Area Under Curve, N = 145

Sample	Variable	AUC	Brier
1	Customer Relations	0.64	0.24

Table B.137: Web Developers
Area Under Curve, N = 192

Sample	Variable	AUC	Brier
1	FIWO*	0.62	0.24

Note B.34. $*$ *Frequent Interaction With Others*

B.3 Model Results with Holdout Samples

Table B.138: Accountants & Auditors Sample
Area Under Curve, N = 981

Sample	Variable	AUC	Brier
1	Consistency	0.60	0.28
2	Nat I	0.56	0.26
3	Consistency	0.61	0.28
4	Nat I	0.61	0.23
5	Nat I	0.59	0.23

Table B.139: Advertising Sales Agents Sample
Area Under Curve, N = 1099

Sample	Variable	AUC	Brier
1	Ad S, FIWO*	0.70	0.21
2	Ad S, FIWO*	0.72	0.20
3	Ad S, FIWO*	0.70	0.21
4	Ad S, FIWO*	0.70	0.20
5	Ad S, Nat I	0.70	0.20

Note B.35. $*$ *Frequent Interaction With Others*

Table B.140: Bookkeeping, Accounting, & Auditing Clerks Sample
Area Under Curve, N = 729

Sample	Variable	AUC	Brier
1	Customer Relations, Ad I	0.67	0.23
2	Versatility, Customer Relations	0.67	0.26
3	Versatility, Customer Relations	0.74	0.20
4	FIWO*, Analysis of Data, Versatility	0.68	0.22
5	Customer Relations, Ad I	0.66	0.22

Note B.36. $*$ *Frequent Interaction With Others*

Table B.141: Chief Executives Sample
Area Under Curve, N = 2370

Sample	Variable	AUC	Brier
1	Versatility*, People Oriented, Consistency, Nat S*	0.70	0.19
2	Ad D, Nat S	0.72	0.19
3	People Oriented, Consistency, Ad D	0.70	0.20
4	People Oriented, FUFT⋆, Ad D	0.71	0.18
5	People Oriented, Frequent Change, Ad D	0.72	0.19

Note B.37. ∗ *Statistical Significance at* 0.90 *level,* ⋆ *Follow Up and Follow Through*

Table B.142: Civil Engineers Sample
Area Under Curve, N = 848

Sample	Variable	AUC	Brier
1	Ad C	0.65	0.26
2	FUFT*	0.59	0.25
3	Ad C	0.61	0.31
4	People Oriented, Versatility	0.61	0.24
5	Ad C	0.60	0.31

Note B.38. ∗ *Follow Up and Follow Through*

Table B.143: Community & Social Service Specialists, All Other
Area Under Curve, N = 1234

Sample	Variable	AUC	Brier
1	Urgency, Nat C	0.61	0.26
2	Ad S	0.58	0.31
3	Ad S	0.64	0.26
4	Ad S	0.60	0.31
5	Ad S	0.60	0.31

Table B.144: Computer Occupations, All Other
Area Under Curve, N = 1037

Sample	Variable	AUC	Brier
1	Ad I	0.58	0.25
2	Ad I	0.56	0.25
3	Ad I	0.59	0.24
4	Ad I, Customer Relations	0.64	0.20
5	Ad I	0.59	0.24

Table B.145: Computer Systems Engineers, Architects
Area Under Curve, N = 586

Sample	Variable	AUC	Brier
1	Nat D, Versatility	0.60	0.21
2	Ad I	0.61	0.23
3	Ad I	0.61	0.25
4	Nat D, Versatility	0.61	0.20
5	Nat D, Versatility	0.60	0.21

Table B.146: Computer User Support Specialists
Area Under Curve, N = 540

Sample	Variable	AUC	Brier
1	Frequent Change	0.63	0.24
2	Nat S	0.63	0.27
3	Ad D	0.65	0.25
4	Following Policy	0.68	0.21
5	Frequent Change	0.66	0.22

Table B.147: Construction & Related Workers, All Other
Area Under Curve, N = 2096

Sample	Variable	AUC	Brier
1	Ad D, FUFT*	0.61	0.21
2	Ad D, FUFT*	0.59	0.22
3	Ad D, FUFT*	0.58	0.23
4	Ad D, Consistency	0.59	0.21
5	Ad D, Consistency	0.61	0.20

Note B.39. $*$ *Follow Up and Follow Through*

Table B.148: Construction Managers
Area Under Curve, N = 1015

Sample	Variable	AUC	Brier
1	Ad D, People Oriented	0.63	0.20
2	People Oriented	0.57	0.26
3	People Oriented	0.59	0.23
4	People Oriented	0.61	0.21
5	Ad D, Customer Relations, People Oriented	0.59	0.23

Table B.149: Customer Service Representatives
Area Under Curve, N = 1232

Sample	Variable	AUC	Brier
1	Nat D, Ad S	0.66	0.26
2	Nat D, People Oriented	0.65	0.24
3	Nat D, Ad S, People Oriented	0.72	0.19
4	Competitiveness	0.66	0.25
5	Ad S, People Oriented	0.65	0.23

Table B.150: Dental Assistants
Area Under Curve, N = 539

Sample	Variable	AUC	Brier
1	Nat I, Ad S	0.69	0.20
2	Ad S, Customer Relations	0.69	0.22
3	Nat I, Ad S	0.76	0.17
4	Ad S, People Oriented	0.73	0.19
5	Nat I, Ad S	0.69	0.20

Table B.151: Dental Hygienists
Area Under Curve, N = 548

Sample	Variable	AUC	Brier
1	Frequent Change, People Oriented	0.75	0.17
2	Ad S, People Oriented	0.65	0.23
3	Frequent Change, People Oriented	0.70	0.20
4	Ad S, People Oriented	0.68	0.22
5	Frequent Change, People Oriented	0.73	0.19

Table B.152: Education Training & Library Workers, All Other
Area Under Curve, N = 1333

Sample	Variable	AUC	Brier
1	Competitiveness	0.56	0.31
2	Customer Relations	0.54	0.33
3	Competitiveness	0.56	0.26
4	Competitiveness*	0.54	0.34
5	Competitiveness, Ad C	0.59	0.22

Note B.40. $*$ *Statistical Significance at* 0.90 *level*

Table B.153: Electricians
Area Under Curve, N = 572

Sample	Variable	AUC	Brier
1	Ad I	0.57	0.25
2	Versatility, Ad D	0.61	0.20
3	Versatility, Nat D	0.65	0.19
4	Versatility, Ad D	0.63	0.21
5	People Oriented, Ad D*	0.66	0.20

Note B.41. $*$ *Statistical Significance at* 0.90 *level*

Table B.154: Executive Secretaries & Executive Administrative Assistants
Area Under Curve, N = 958

Sample	Variable	AUC	Brier
1	Nat D	0.68	0.24
2	Ad S, Customer Relations	0.65	0.24
3	Ad S, Customer Relations	0.65	0.23
4	Competitiveness, Customer Relations	0.65	0.24
5	Ad S, Customer Relations	0.66	0.24

Table B.155: Financial Analysts
Area Under Curve, N = 1317

Sample	Variable	AUC	Brier
1	People Oriented	0.59	0.22
2	*		
3	People Oriented	0.59	0.22
4	People Oriented	0.54	0.25
5	People Oriented, Nat D	0.59	0.23

Note B.42. * *No statistically significant model was found for this sample*

Table B.156: Financial Specialists, All Other
Area Under Curve, N = 2072

Sample	Variable	AUC	Brier
1	Urgency	0.53	0.26
2	Ad S	0.56	0.24
3	Ad S	0.53	0.26
4	Nat S	0.55	0.26
5	*		

Note B.43. * *No statistically significant model was found for this sample*

Table B.157: General & Operations Managers
Area Under Curve, N = 2316

Sample	Variable	AUC	Brier
1	Nat S, FC*	0.60	0.22
2	Nat S, FC*	0.60	0.21
3	Nat S, FC*	0.59	0.22
4	Nat S, FC*	0.59	0.22
5	Nat S, FC*	0.60	0.22

Note B.44. * *Frequent Change*

Table B.158: Healthcare & Support Workers, All Other
Area Under Curve, N = 2532

Sample	Variable	AUC	Brier
1	Ad S, CR*	0.60	0.26
2	Ad S, CR*	0.60	0.25
3	Ad S	0.58	0.31
4	CR*	0.59	0.26
5	Ad S	0.57	0.31

Note B.45. ∗ *Customer Relations*

Table B.159: Human Resource Assistants, Except Payroll & Timekeeping
Area Under Curve, N = 782

Sample	Variable	AUC	Brier
1	FIWO, Comp*	0.67	0.21
2	Comp*	0.65	0.28
3	People Oriented	0.59	0.26
4	FIWO, Comp*	0.66	0.21
5	FIWO, Comp*	0.69	0.20

Note B.46. ∗ *Frequent Interaction With Others, Competitiveness*

Table B.160: Human Resources Managers
Area Under Curve, N = 1642

Sample	Variable	AUC	Brier
1	FIWO*	0.54	0.34
2	FIWO*	0.58	0.31
3	FUFT⋆	0.55	0.27
4	FUFT⋆	0.58	0.24
5	FUFT⋆	0.56	0.27

Note B.47. ∗ *Frequent Interaction With Others,* ⋆ *Follow Up and Follow Through*

Table B.161: Human Resources Specialists
Area Under Curve, N = 1307

Sample	Variable	AUC	Brier
1	*		
2	FUFT⋆	0.58	0.31
3	FUFT⋆	0.55	0.27
4	FUFT⋆	0.58	0.24
5	*		

Note B.48. ∗ *No significant model for this data set,* ⋆ *Follow Up and Follow Through*

Table B.162: Information Technology Project Managers
Area Under Curve, N = 747

Sample	Variable	AUC	Brier
1	Ad C	0.57	0.31
2	Ad C	0.62	0.29
3	People Oriented	0.57	0.23
4	*		
5	*		

Note B.49. ∗ *No significant model for this data set.*

Table B.163: Insurance Sales Agents
Area Under Curve, N = 968

Sample	Variable	AUC	Brier
1	Consistency	0.61	0.23
2	Consistency, Nat I	0.67	0.23
3	Consistency	0.63	0.22
4	Consistency, Nat I	0.64	0.23
5	Consistency	0.61	0.22

Table B.164: Lawyers
Area Under Curve, N = 1534

Sample	Variable	AUC	Brier
1	FIWO*	0.56	0.26
2	Ad C	0.57	0.31
3	Ad C	0.55	0.33
4	Ad C	0.58	0.31
5	Ad C	0.59	0.32

Note B.50. ∗ *Frequent Interaction With Others*

Table B.165: Management Analysts
Area Under Curve, N = 512

Sample	Variable	AUC	Brier
1	Competitiveness	0.58	0.26
2	Competitiveness	0.57	0.26
3	∗		
4	Competitiveness	0.59	0.26
5	Competitiveness	0.60	0.25

Note B.51. ∗ *No significant model for this data set.*

Table B.166: Managers, All Other
Area Under Curve, N = 4199

Sample	Variable	AUC	Brier
1	CR⋆	0.56	0.23
2	CR⋆	0.53	0.24
3	∗		
4	AoD†	0.53	0.31
5	CR⋆	0.53	0.24

Note B.52. ∗ *No significant model for this data set,* ⋆ *Customer Relations,* † *Analysis of Data.*

Table B.167: Market Research Analysts & Marketing Specialists
Area Under Curve, N = 1503

Sample	Variable	AUC	Brier
1	Ad I	0.57	0.32
2	Ad I	0.56	0.32
3	Ad I	0.59	0.30
4	Ad I	0.59	0.30
5	Ad I	0.56	0.33

Table B.168: Marketing Managers
Area Under Curve, N = 1424

Sample	Variable	AUC	Brier
1	FUFT*	0.61	0.24
2	FUFT*	0.59	0.25
3	FUFT*	0.63	0.23
4	FUFT*	0.58	0.26
5	FUFT*	0.64	0.22

Note B.53. ∗ *Follow Up and Follow Through*

Table B.169: Mechanical Engineers
Area Under Curve, N = 695

Sample	Variable	AUC	Brier
1	PO*, Nat C	0.69	0.20
2	PO*, Nat C	0.63	0.21
3	AoD⋆	0.61	0.27
4	PO*, FIWO†	0.63	0.21
5	AoD⋆	0.66	0.24

Note B.54. ∗ *People Oriented,* ⋆ *Analysis of Data,* † *Frequent Interaction With Others*

Table B.170: Media & Communication Workers, All Other
Area Under Curve, N = 815

Sample	Variable	AUC	Brier
1	FIWO*, Nat C	0.59	0.32
2	FIWO*, FP⋆, †	0.59	0.24
3	FIWO*, FP⋆	0.60	0.23
4	FIWO*	0.58	0.32
5	FIWO*, FP⋆	0.63	0.22

Note B.55. ∗ *Frequent Interaction With Others,* ⋆ *Following Policy,* † *Model is significant at the* 90% *confidence level.*

Table B.171: Office & Administrative Support Workers, All Other
Area Under Curve, N = 2559

Sample	Variable	AUC	Brier
1	PO*, Ad S, Comp⋆	0.68	0.22
2	PO*, Ad S, Nat D	0.69	0.21
3	PO*, Urg†	0.66	0.27
4	PO*, Ad S, Nat D	0.69	0.22
5	PO*, Ad S, Nat D, ††	0.69	0.22

Note B.56. ∗ *People Oriented,* ⋆ *Competitiveness,* † *Urgency,* †† *Model is significant at the* 90% *confidence level.*

Table B.172: Paralegals & Legal Assistants
Area Under Curve, N = 505

Sample	Variable	AUC	Brier
1	CR*, Ad S	0.65	0.23
2	Ad S	0.63	0.27
3	Ad S	0.65	0.25
4	Ad S	0.63	0.27
5	Ad S	0.63	0.27

Note B.57. ∗ *Customer Relations*

Table B.173: Production Workers, All Other
Area Under Curve, N = 500

Sample	Variable	AUC	Brier
1	Analysis of Data	0.59	0.28
2	Nat S	0.59	0.29
3	Nat S	0.58	0.32
4	Nat S*	0.54	0.33
5	Nat S*	0.55	0.33

Note B.58. $*$ *Models significant at the* 90% *confidence level.*

Table B.174: Real Estate Sales Agents
Area Under Curve, N = 889

Sample	Variable	AUC	Brier
1	Nat I	0.68	0.23
2	Nat I, Nat C	0.65	0.27
3	Nat I, Nat C	0.69	0.22
4	Nat C	0.69	0.21
5	Nat C	0.68	0.22

Table B.175: Registered Nurses
Area Under Curve, N = 1099

Sample	Variable	AUC	Brier
1	Comp, PO*	0.67	0.22
2	Ad S	0.59	0.31
3	Ad S	0.60	0.28
4	Ad S	0.62	0.29
5	Comp	0.60	0.31

Note B.59. $*$ *Competitiveness, People Oriented*

Table B.176: Sales Agents, Financial Services
Area Under Curve, N = 822

Sample	Variable	AUC	Brier
1	Consistency	0.69	0.20
2	Consistency	0.64	0.22
3	Consistency	0.64	0.23
4	Versatility	0.68	0.22
5	Versatility	0.67	0.22

Table B.177: Sales & Related Workers, All Other
Area Under Curve, N = 2919

Sample	Variable	AUC	Brier
1	FC*, Ad I	0.64	0.23
2	FC*, Ad I	0.64	0.23
3	FC*, OW†	0.66	0.24
4	FC*, OW†	0.63	0.25
5	FC*	0.64	0.22

Note B.60. ∗ *Frequent Change,* † *Organized Workplace*

Table B.178: Sales Engineers
Area Under Curve, N = 630

Sample	Variable	AUC	Brier
1	Versatility	0.58	0.27
2	Versatility	0.61	0.25
3	Following Policy	0.63	0.24
4	Versatility	0.59	0.25
5	Following Policy	0.63	0.23

Table B.179: Sales Managers
Area Under Curve, N = 1224

Sample	Variable	AUC	Brier
1	Ad S	0.68	0.21
2	Versatility, Ad S	0.69	0.21
3	Ad S	0.68	0.21
4	Versatility, Ad D	0.69	0.20
5	Versatility, Ad D	0.68	0.20

Table B.180: Sales Representatives, Services, All Other
Area Under Curve, N = 7749

Sample	Variable	AUC	Brier
1	AoD*, Ad C, Nat S	0.66	0.21
2	AoD*, Ad C, Nat S	0.67	0.21
3	AoD*, Cons†	0.67	0.22
4	AoD*, FP⋆	0.67	0.22
5	AoD*, Cons†	0.67	0.22

Note B.61. ∗ *Analysis of Data,* † *Consistency,* ⋆ *Following Policy*

Table B.181: Sales Representatives, Wholesale & Manufacturing, Except Technical & Science
Area Under Curve, N = 1419

Sample	Variable	AUC	Brier
1	AoD*, Ad S	0.70	0.20
2	OW†, Ad S	0.63	0.25
3	OW†, Ad S	0.64	0.25
4	OW†, Ad S	0.65	0.25
5	OW†, Ad S	0.66	0.24

Note B.62. ∗ *Analysis of Data,* † *Organized Workplace*

Table B.182: Sales Representatives, Wholesale & Manufacturing, Technical & Science
Area Under Curve, N = 1966

Sample	Variable	AUC	Brier
1	Consistency	0.69	0.20
2	Frequent Change	0.66	0.21
3	Frequent Change	0.66	0.21
4	Versatility, Nat D	0.67	0.20
5	Versatility, Nat D	0.67	0.20

Table B.183: Software Developers, Applications
Area Under Curve, N = 620

Sample	Variable	AUC	Brier
1	Nat I	0.61	0.23
2	Nat I	0.61	0.24
3	Nat I	0.59	0.23
4	Nat I	0.64	0.22
5	Nat I, Consistency*	0.64	0.22

Note B.63. $*$ *model significant at the* 90% *confidence level for this data set.*

Table B.184: Transportation Workers, All Other
Area Under Curve, N = 650

Sample	Variable	AUC	Brier
1	People Oriented	0.57	0.24
2	People Oriented	0.59	0.23
3	People Oriented	0.56	0.24
4	Versatility	0.59	0.24
5	Versatility*	0.55	0.25

Note B.64. $*$ *model significant at the* 90% *confidence level for this data set.*

C Historical Item Difficulty and Discrimination Charts

This section contains the historical item analysis charts for the TTI Success Insights Style Insights assessment over the last decade. The current item analysis results are presented in Section 4.2. For a more detailed discussion of item analysis, one may read Section 4.1. One may also consult Chapter 7 of [74] for a more complete discussion on the relationship between item analysis, internal structure validity, and internal consistency.

C.1 2012 Historical Item Analysis

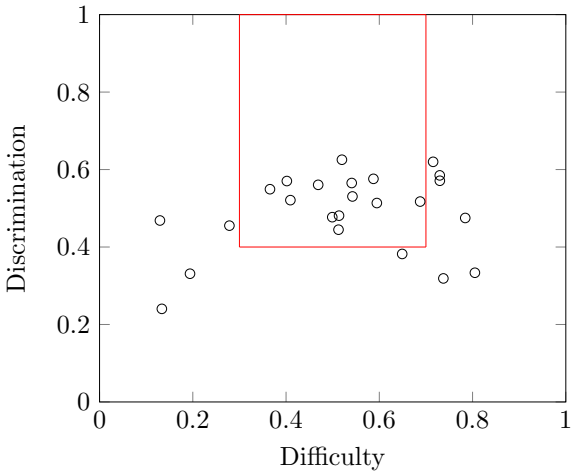

Figure C.1: 2012 Forced-rank Dominance

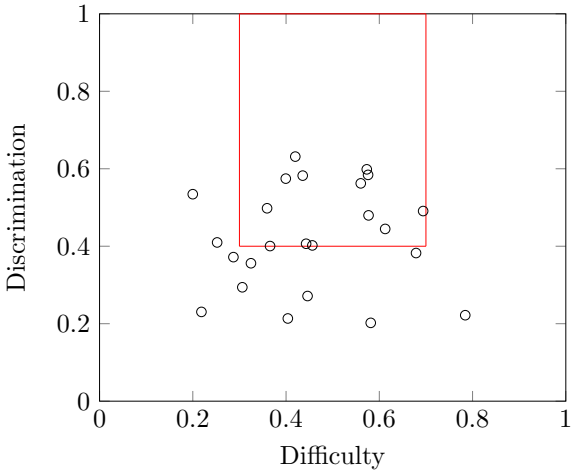

Figure C.2: 2012 Forced-rank Influence

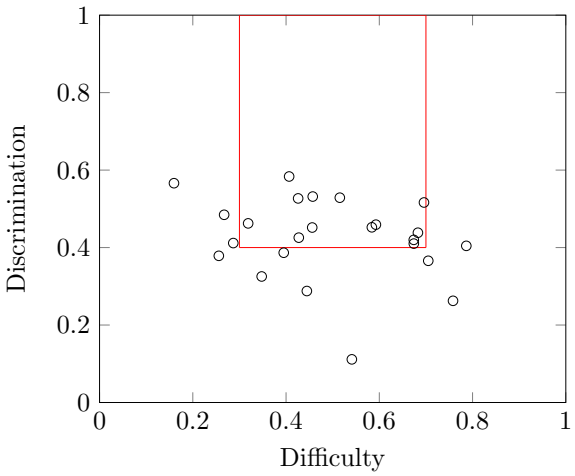

Figure C.3: 2012 Forced-rank Steadiness

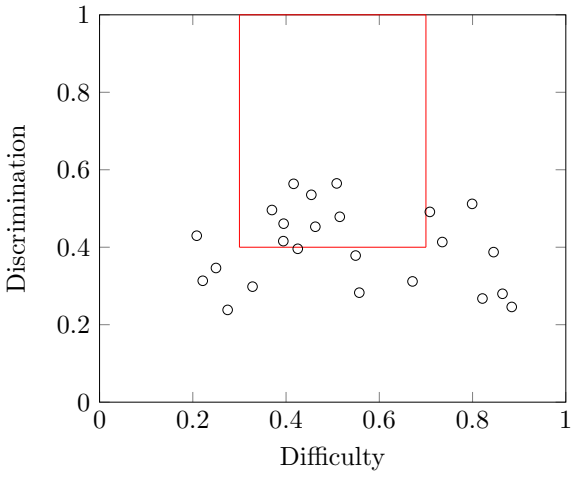

Figure C.4: 2012 Forced-rank Compliance

C.2 2013 Historical Item Analysis

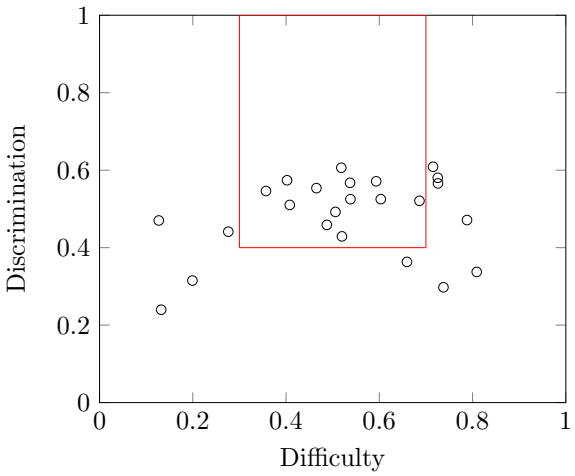

Figure C.5: 2013 Forced-rank Dominance

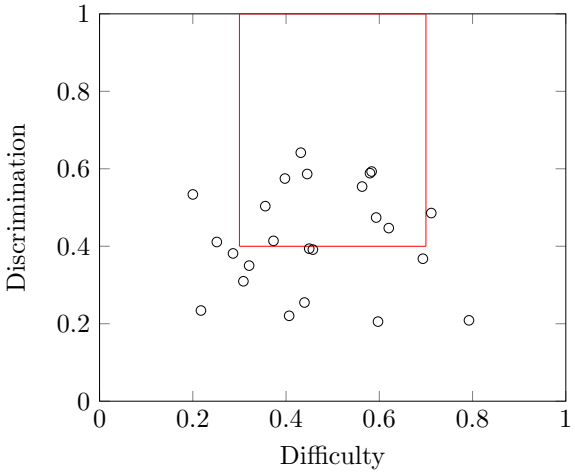

Figure C.6: 2013 Forced-rank Influence

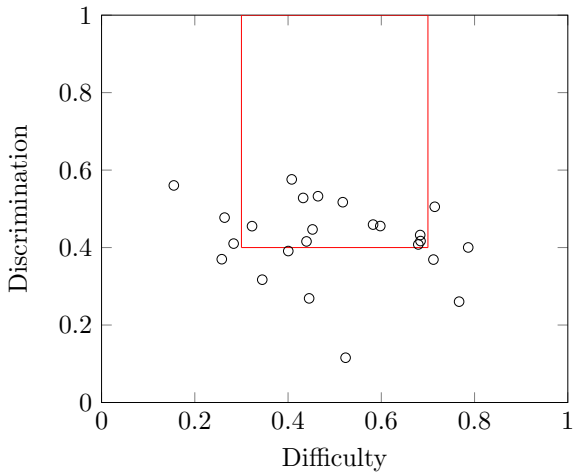

Figure C.7: 2013 Forced-rank Steadiness

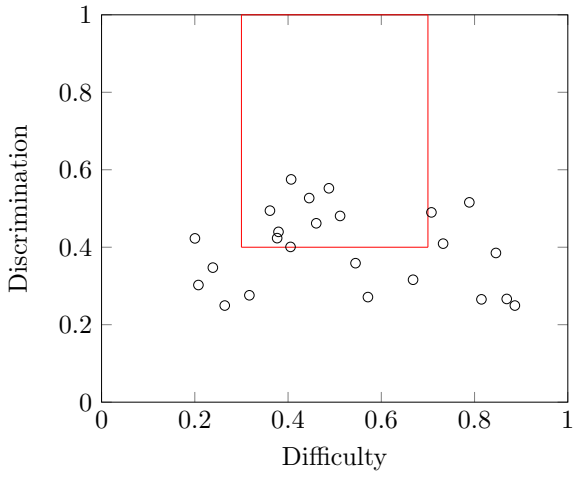

Figure C.8: 2013 Forced-rank Compliance

C.3 2017 Historical Item Analysis

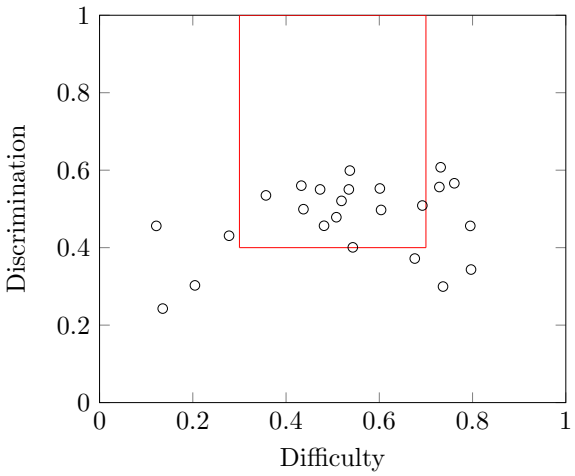

Figure C.9: 2017 Forced-rank Dominance

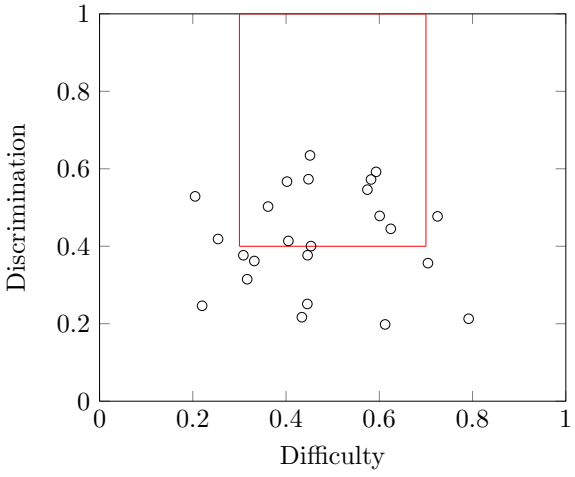

Figure C.10: 2017 Forced-rank Influence

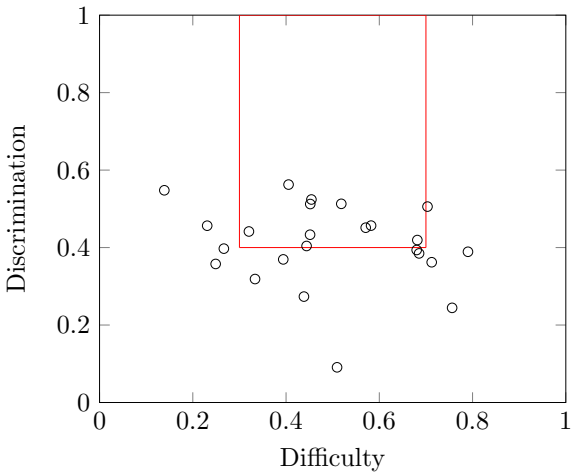

Figure C.11: 2017 Forced-rank Steadiness

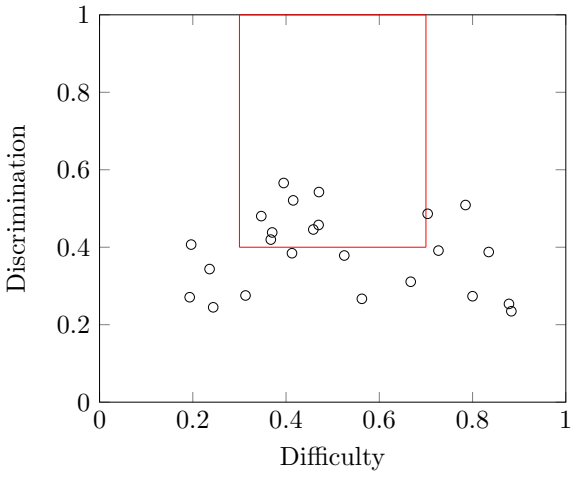

Figure C.12: 2017 Forced-rank Compliance

D Historical Inter-item Correlation Charts

This section contains the historical inter-item correlation charts for the TTI Success Insights Style Insights assessment over the last decade. The current inter-item correlation results are presented in Section 5.2. For a more detailed discussion of inter-item correlation, one may read Section 5.1. One may also consult Chapter 4 of [74] for a more complete discussion on the relationship between inter-item correlation and internal consistency.

D.1 2012 Historical Inter-item Correlation

Table D.1: 2012 Average English US Inter-item Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.28	0.21	0.22	0.19
Item 2	0.26	0.21	0.21	0.18
Item 3	0.26	0.21	0.21	0.19
Item 4	0.26	0.22	0.21	0.19
Item 5	0.27	0.20	0.20	0.18
Item 6	0.27	0.20	0.21	0.19
Item 7	0.26	0.21	0.21	0.19
Item 8	0.27	0.20	0.21	0.18
Item 9	0.26	0.20	0.21	0.19
Item 10	0.27	0.20	0.21	0.19
Item 11	0.26	0.21	0.21	0.19
Item 12	0.27	0.20	0.21	0.18
Item 13	0.27	0.21	0.21	0.18
Item 14	0.27	0.20	0.22	0.19
Item 15	0.26	0.20	0.20	0.18
Item 16	0.26	0.21	0.20	0.19
Item 17	0.26	0.21	0.21	0.18
Item 18	0.26	0.20	0.20	0.19
Item 19	0.27	0.20	0.20	0.18
Item 20	0.27	0.21	0.21	0.19
Item 21	0.26	0.21	0.21	0.19
Item 22	0.26	0.21	0.21	0.18
Item 23	0.26	0.22	0.21	0.19
Item 24	0.27	0.21	0.21	0.19
AVG	0.27	0.21	0.21	0.19

D.2 2013 Historical Inter-item Correlation

Table D.2: 2013 Average English US Inter-item Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.27	0.21	0.21	0.19
Item 2	0.25	0.20	0.21	0.18
Item 3	0.26	0.21	0.20	0.19
Item 4	0.26	0.21	0.20	0.19
Item 5	0.27	0.20	0.20	0.18
Item 6	0.26	0.20	0.20	0.18
Item 7	0.26	0.21	0.21	0.18
Item 8	0.26	0.20	0.21	0.18
Item 9	0.26	0.20	0.21	0.19
Item 10	0.26	0.20	0.21	0.19
Item 11	0.26	0.21	0.20	0.19
Item 12	0.26	0.20	0.21	0.18
Item 13	0.26	0.21	0.21	0.18
Item 14	0.27	0.20	0.22	0.18
Item 15	0.26	0.20	0.20	0.18
Item 16	0.25	0.21	0.20	0.19
Item 17	0.26	0.21	0.21	0.18
Item 18	0.26	0.20	0.20	0.18
Item 19	0.27	0.20	0.20	0.18
Item 20	0.27	0.21	0.21	0.19
Item 21	0.26	0.21	0.20	0.19
Item 22	0.26	0.21	0.21	0.18
Item 23	0.26	0.21	0.20	0.18
Item 24	0.26	0.21	0.21	0.19
AVG	0.26	0.21	0.21	0.18

D.3 2017 Historical Inter-item Correlation

Table D.3: 2017 Average English US Inter-item Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.26	0.21	0.20	0.18
Item 2	0.24	0.20	0.20	0.17
Item 3	0.25	0.21	0.19	0.18
Item 4	0.25	0.21	0.19	0.18
Item 5	0.26	0.20	0.19	0.17
Item 6	0.25	0.20	0.19	0.18
Item 7	0.25	0.21	0.20	0.18
Item 8	0.25	0.20	0.20	0.17
Item 9	0.25	0.20	0.20	0.18
Item 10	0.25	0.20	0.20	0.18
Item 11	0.25	0.21	0.19	0.18
Item 12	0.25	0.20	0.20	0.18
Item 13	0.25	0.21	0.20	0.18
Item 14	0.26	0.19	0.21	0.18
Item 15	0.25	0.20	0.19	0.17
Item 16	0.24	0.21	0.19	0.18
Item 17	0.25	0.21	0.20	0.17
Item 18	0.25	0.20	0.19	0.18
Item 19	0.26	0.20	0.19	0.17
Item 20	0.26	0.20	0.20	0.18
Item 21	0.25	0.21	0.19	0.18
Item 22	0.25	0.20	0.20	0.17
Item 23	0.25	0.21	0.19	0.18
Item 24	0.25	0.20	0.20	0.18
AVG	0.25	0.20	0.20	0.18

E Historical Item-Total CorrelationCharts

This section contains the historical corrected item-total correlation charts for the TTI Success Insights Style Insights assessment over the last decade. The current corrected item-total correlation results are presented in Section 6.2. For a more detailed discussion of corrected item-total correlation, one may read Section 6.1. One may also consult Chapter 7 of [74] for a more complete discussion on the relationship between corrected item-total correlation and internal structure validity.

E.1 2012 Historical Corrected Item-total Correlation

Table E.1: 2012 English US Corrected Item-total Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.24	0.37	0.26	0.27
Item 2	0.62	0.44	0.38	0.56
Item 3	0.57	0.36	0.52	0.35
Item 4	0.52	0.20	0.45	0.24
Item 5	0.38	0.58	0.58	0.56
Item 6	0.48	0.49	0.45	0.43
Item 7	0.51	0.29	0.37	0.42
Item 8	0.44	0.50	0.41	0.54
Item 9	0.56	0.60	0.41	0.31
Item 10	0.47	0.53	0.40	0.25
Item 11	0.58	0.27	0.48	0.28
Item 12	0.46	0.58	0.44	0.45
Item 13	0.48	0.38	0.43	0.46
Item 14	0.33	0.63	0.11	0.39
Item 15	0.52	0.56	0.53	0.48
Item 16	0.63	0.40	0.57	0.28
Item 17	0.58	0.23	0.39	0.51
Item 18	0.57	0.48	0.53	0.40
Item 19	0.33	0.57	0.53	0.50
Item 20	0.32	0.40	0.33	0.38
Item 21	0.57	0.22	0.46	0.31
Item 22	0.55	0.41	0.42	0.49
Item 23	0.53	0.21	0.46	0.41
Item 24	0.47	0.41	0.29	0.30
AVG	0.49	0.42	0.42	0.40
Std.Dev.	0.57	0.50	0.48	0.48

E.2 2013 Historical Corrected Item-total Correlation

Table E.2: 2013 English US Corrected Item-total Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.24	0.38	0.26	0.27
Item 2	0.61	0.45	0.37	0.58
Item 3	0.57	0.35	0.51	0.35
Item 4	0.52	0.21	0.45	0.25
Item 5	0.36	0.59	0.58	0.55
Item 6	0.49	0.49	0.46	0.42
Item 7	0.53	0.31	0.37	0.42
Item 8	0.43	0.50	0.41	0.53
Item 9	0.55	0.59	0.41	0.32
Item 10	0.47	0.53	0.40	0.25
Item 11	0.58	0.25	0.48	0.27
Item 12	0.44	0.59	0.43	0.46
Item 13	0.46	0.37	0.42	0.44
Item 14	0.32	0.64	0.12	0.39
Item 15	0.51	0.55	0.52	0.48
Item 16	0.61	0.39	0.56	0.27
Item 17	0.57	0.23	0.39	0.52
Item 18	0.57	0.47	0.53	0.40
Item 19	0.34	0.57	0.53	0.49
Item 20	0.30	0.41	0.32	0.36
Item 21	0.57	0.21	0.46	0.30
Item 22	0.55	0.41	0.42	0.49
Item 23	0.53	0.22	0.46	0.41
Item 24	0.47	0.39	0.27	0.28
AVG	0.48	0.42	0.42	0.40
Std.Dev.	0.57	0.50	0.48	0.48

E.3 2017 Historical Corrected Item-total Correlation

Table E.3: 2017 English US Corrected Item-total Correlations
TTI SI Style Insights

Item/Scale	D	I	S	C
Item 1	0.24	0.38	0.24	0.27
Item 2	0.61	0.45	0.36	0.57
Item 3	0.56	0.36	0.51	0.34
Item 4	0.51	0.20	0.43	0.24
Item 5	0.37	0.57	0.56	0.54
Item 6	0.48	0.48	0.45	0.41
Item 7	0.50	0.31	0.36	0.42
Item 8	0.40	0.50	0.39	0.52
Item 9	0.55	0.59	0.40	0.31
Item 10	0.46	0.53	0.39	0.23
Item 11	0.57	0.25	0.46	0.27
Item 12	0.43	0.57	0.42	0.45
Item 13	0.46	0.36	0.40	0.44
Item 14	0.30	0.63	0.09	0.39
Item 15	0.50	0.55	0.51	0.46
Item 16	0.60	0.38	0.55	0.25
Item 17	0.55	0.25	0.37	0.51
Item 18	0.55	0.48	0.51	0.38
Item 19	0.34	0.57	0.52	0.48
Item 20	0.30	0.41	0.32	0.38
Item 21	0.56	0.21	0.46	0.27
Item 22	0.53	0.42	0.38	0.49
Item 23	0.52	0.22	0.44	0.39
Item 24	0.46	0.40	0.27	0.28
AVG	0.47	0.42	0.41	0.39
Std.Dev.	0.57	0.50	0.48	0.48

F Internal Consistency Tables

This section contains the historical internal consistency charts for the TTI Success Insights Style Insights assessment over the last decade. The current internal consistency results are presented in Section 7.2.1. For a more detailed discussion of internal consistency, one may read Sections 7.1 and 7.2. One may also consult Chapter 4 of [74] for a more complete discussion on the relationship between internal consistency and reliability.

F.1 2012 Historical Internal Consistency

Table F.1: 2012 Average English US α Coefficient Data
TTI SI Style Insights

Est./Scale	D	I	S	C
α	0.897	0.862	0.864	0.847
S.E.	0.001	0.002	0.002	0.002
Upper CI	0.900	0.865	0.867	0.850
Lower CI	0.893	0.859	0.861	0.843

F.2 2013 Historical Internal Consistency

Table F.2: 2013 Average English US α Coefficient Data
TTI SI Style Insights

Est./Scale	D	I	S	C
α	0.894	0.861	0.861	0.844
S.E.	0.001	0.001	0.001	0.002
Upper CI	0.897	0.864	0.864	0.847
Lower CI	0.891	0.858	0.858	0.841

F.3 2017 Historical Internal Consistency

Table F.3: 2017 Average English US α Coefficient Data
TTI SI Style Insights

Est./Scale	D	I	S	C
α	0.889	0.860	0.854	0.839
S.E.	0.001	0.001	0.001	0.002
Upper CI	0.892	0.863	0.857	0.842
Lower CI	0.886	0.857	0.851	0.835

G α-if-item-deleted Tables

This section contains the historical α-if-item-deleted charts for the TTI Success Insights Style Insights assessment over the last decade. The current α-if-item-deleted results are presented in Section 7.4.

G.1 2012 Historical α-if-item-deleted

Table G.1: 2012 English US α-if-Item-Deleted Data
Dominance,$\alpha = 0.897$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.898	Item 7	0.892	Item 13	0.893	Item 19	0.896
Item 2	0.889	Item 8	0.894	Item 14	0.896	Item 20	0.897
Item 3	0.891	Item 9	0.891	Item 15	0.892	Item 21	0.891
Item 4	0.892	Item 10	0.893	Item 16	0.889	Item 22	0.891
Item 5	0.895	Item 11	0.890	Item 17	0.890	Item 23	0.891
Item 6	0.893	Item 12	0.893	Item 18	0.891	Item 24	0.893

Table G.2: 2012 English US α-if-Item-Deleted Data
Influence,$\alpha = 0.862$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.858	Item 7	0.861	Item 13	0.858	Item 19	0.851
Item 2	0.856	Item 8	0.854	Item 14	0.850	Item 20	0.857
Item 3	0.859	Item 9	0.851	Item 15	0.852	Item 21	0.863
Item 4	0.863	Item 10	0.853	Item 16	0.857	Item 22	0.857
Item 5	0.851	Item 11	0.861	Item 17	0.863	Item 23	0.863
Item 6	0.854	Item 12	0.851	Item 18	0.855	Item 24	0.857

Table G.3: 2012 English US α-if-Item-Deleted Data
Steadiness,$\alpha = 0.864$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.864	Item 7	0.861	Item 13	0.859	Item 19	0.855
Item 2	0.860	Item 8	0.859	Item 14	0.868	Item 20	0.862
Item 3	0.856	Item 9	0.859	Item 15	0.855	Item 21	0.858
Item 4	0.858	Item 10	0.859	Item 16	0.854	Item 22	0.859
Item 5	0.854	Item 11	0.857	Item 17	0.860	Item 23	0.857
Item 6	0.858	Item 12	0.858	Item 18	0.855	Item 24	0.863

Table G.4: 2012 English US α-if-Item-Deleted Data
Compliance,$\alpha = 0.847$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.845	Item 7	0.841	Item 13	0.839	Item 19	0.838
Item 2	0.835	Item 8	0.836	Item 14	0.841	Item 20	0.842
Item 3	0.843	Item 9	0.844	Item 15	0.838	Item 21	0.844
Item 4	0.847	Item 10	0.846	Item 16	0.845	Item 22	0.837
Item 5	0.835	Item 11	0.845	Item 17	0.837	Item 23	0.840
Item 6	0.840	Item 12	0.839	Item 18	0.841	Item 24	0.845

G.2 2013 Historical α-if-item-deleted

Table G.5: 2013 English US α-if-Item-Deleted Data
Dominance,$\alpha = 0.894$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.895	Item 7	0.889	Item 13	0.890	Item 19	0.893
Item 2	0.887	Item 8	0.891	Item 14	0.894	Item 20	0.894
Item 3	0.888	Item 9	0.888	Item 15	0.889	Item 21	0.887
Item 4	0.889	Item 10	0.890	Item 16	0.887	Item 22	0.888
Item 5	0.893	Item 11	0.887	Item 17	0.888	Item 23	0.889
Item 6	0.889	Item 12	0.890	Item 18	0.888	Item 24	0.890

Table G.6: 2013 English US α-if-Item-Deleted Data
Influence,$\alpha = 0.861$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.857	Item 7	0.860	Item 13	0.858	Item 19	0.851
Item 2	0.855	Item 8	0.853	Item 14	0.849	Item 20	0.856
Item 3	0.858	Item 9	0.850	Item 15	0.852	Item 21	0.863
Item 4	0.863	Item 10	0.852	Item 16	0.857	Item 22	0.856
Item 5	0.851	Item 11	0.861	Item 17	0.862	Item 23	0.862
Item 6	0.854	Item 12	0.851	Item 18	0.854	Item 24	0.857

Table G.7: 2013 English US α-if-Item-Deleted Data
Steadiness,$\alpha = 0.861$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.861	Item 7	0.858	Item 13	0.856	Item 19	0.853
Item 2	0.858	Item 8	0.856	Item 14	0.866	Item 20	0.859
Item 3	0.853	Item 9	0.856	Item 15	0.853	Item 21	0.855
Item 4	0.855	Item 10	0.857	Item 16	0.851	Item 22	0.856
Item 5	0.851	Item 11	0.854	Item 17	0.857	Item 23	0.855
Item 6	0.855	Item 12	0.856	Item 18	0.852	Item 24	0.861

Table G.8: 2013 English US α-if-Item-Deleted Data
Compliance,$\alpha = 0.844$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.843	Item 7	0.838	Item 13	0.837	Item 19	0.835
Item 2	0.832	Item 8	0.834	Item 14	0.838	Item 20	0.840
Item 3	0.840	Item 9	0.841	Item 15	0.835	Item 21	0.842
Item 4	0.844	Item 10	0.843	Item 16	0.843	Item 22	0.834
Item 5	0.833	Item 11	0.843	Item 17	0.834	Item 23	0.837
Item 6	0.838	Item 12	0.836	Item 18	0.839	Item 24	0.843

G.3 2017 Historical α-if-item-deleted

Table G.9: 2017 English US α-if-Item-Deleted Data
Dominance,$\alpha = 0.889$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.891	Item 7	0.884	Item 13	0.885	Item 19	0.888
Item 2	0.882	Item 8	0.887	Item 14	0.889	Item 20	0.889
Item 3	0.883	Item 9	0.883	Item 15	0.884	Item 21	0.883
Item 4	0.884	Item 10	0.885	Item 16	0.882	Item 22	0.883
Item 5	0.888	Item 11	0.883	Item 17	0.883	Item 23	0.884
Item 6	0.885	Item 12	0.886	Item 18	0.883	Item 24	0.886

Table G.10: 2017 English US α-if-Item-Deleted Data
Influence,$\alpha = 0.860$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.856	Item 7	0.858	Item 13	0.857	Item 19	0.849
Item 2	0.853	Item 8	0.852	Item 14	0.847	Item 20	0.855
Item 3	0.856	Item 9	0.849	Item 15	0.850	Item 21	0.861
Item 4	0.862	Item 10	0.851	Item 16	0.856	Item 22	0.854
Item 5	0.849	Item 11	0.860	Item 17	0.860	Item 23	0.861
Item 6	0.853	Item 12	0.849	Item 18	0.852	Item 24	0.855

Table G.11: 2017 English US α-if-Item-Deleted Data
Steadiness, $\alpha = 0.854$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.854	Item 7	0.850	Item 13	0.849	Item 19	0.845
Item 2	0.850	Item 8	0.849	Item 14	0.859	Item 20	0.852
Item 3	0.845	Item 9	0.849	Item 15	0.845	Item 21	0.847
Item 4	0.848	Item 10	0.849	Item 16	0.844	Item 22	0.850
Item 5	0.843	Item 11	0.847	Item 17	0.850	Item 23	0.847
Item 6	0.847	Item 12	0.848	Item 18	0.845	Item 24	0.853

Table G.12: 2017 English US α-if-Item-Deleted Data
Compliance,$\alpha = 0.839$

Item/Value	α	Item/Value	α	Item/Value	α	Item/Value	α
Item 1	0.837	Item 7	0.832	Item 13	0.831	Item 19	0.830
Item 2	0.826	Item 8	0.828	Item 14	0.832	Item 20	0.833
Item 3	0.834	Item 9	0.835	Item 15	0.830	Item 21	0.838
Item 4	0.838	Item 10	0.838	Item 16	0.837	Item 22	0.828
Item 5	0.827	Item 11	0.837	Item 17	0.828	Item 23	0.832
Item 6	0.832	Item 12	0.830	Item 18	0.833	Item 24	0.837

H Assessment Development Outline