TTI Success Insights' Approach to Psychometric Assessment Validity and Reliability

Eric Gehrig, PhD*
Senior Research Scientist
Research & Development
Target Training International

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

May 1, 2019

1 Introduction

As noted in a previous article published in the April edition of the TTI Master Distributor monthly international newsletter, see *What is Happening at TTI SI*, the concepts of establishing evidence of validity and reliability are more complex and challenging than ever before. There is clearly a general consensus surrounding the need for psychometric assessment validation. However, the literature suggests a disconnect between the stated desires of the academic community and the application by practitioners. We begin with a short example of the argument supporting the previous claim and then move on to the approach that TTI Success Insights has in place comply with the desires of the wider psychometric community.

The example is related to evidence of validity measures related to factor analytic techniques that may be used to measure the uni-dimensionality and structural validity evidence for an assessment. Without detailing all aspects of factor analytic approaches use assessment validation, a widely used metric is that of variance explained, variance accounted for, or total variance accounted for. The terms are synonymous and used in slightly different application of reporting factor analytic results.

One main question to ask and answer is what is an acceptable level of variance explained. This is where the inconsistencies come in. There are widely varying descriptions of both suggested approaches to applying factor analysis and acceptable levels of measurement of variance explained. A classic work on the topic, [29], notes that a usual application of practitioners is to factor analyze an assessment scale until "75, 80, or 85% of the variance is accounted for." However, in the next paragraph of the same work the author states:

> A predetermined level of variance to be extracted may not be set; a table of the percentages may be examined instead. Usually factor extraction is stopped after a large portion of the variance has been extracted and when the next factor would add only a very small percentage to the total variance extracted.

This author appears to be stating that the better approach to setting a predetermined level of variance accounted for to achieve is to measure the effectiveness of the factor analysis as part of a

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

marginal cost analysis. If the benefit outweighs the cost, continue factor analysis, otherwise stop and the variance explained is measured to be what it is. This is not to say that any level of variance explained is acceptable. As noted in [31],

> ... the practical realization of explaining 70% or better of the matrix of association variance is a rarity in much of the published literature. Nevertheless, many EFAs in the present study failed to account for even a reasonable amount of variance (i.e., 28.6% of articles reporting variance accounted for explained less than 30% of the matrix variance).

It is not clear that the authors of this paper are suggesting that 30% variance accounted for is a reasonable level to report at or above. However, one may draw some inferences from other information contained in the same report. For example, this paper reports a meta-analysis of 109 papers published in this area with a mean reported variance accounted for of 45% with a standard deviation of 17%. Assuming a normally distributed set of data on the topic, reporting variance explained at or above 30% is within a standard deviation of the mean. This may be a reasonable, although not desired, level of variance explained to report. Further reading shows that the range of variance explained spanned a low of 12.8% to a high of 70.2%. This is quite a large range.

As a final viewpoint, we present that of the authors of [36].

> While there is no universal standard available to make this decision, the idea is to use only those variables that have a strong association with the factor. A threshold value that can be used for structure or pattern coefficients is equal to or greater than a magnitude of 0.40.

Note that the value referred to in the quote is not the variance explained but the actual factor load of an item on a factor. The "universal standard" referred to in the quote is the variance explained value. It appears that the agreement between the references cited here is that there is no generally acceptable minimum level of variance explained, that factor analysis should be conducted with a marginal cost mindset, and a minimal value for a factor load is a magnitude of 0.40.

To conclude the introduction to this paper, many examples of inconsistencies exist related to acceptable levels of validity or internal consistency and so on. The main reason that inconsistencies exist is that the full picture of an assessment, all its uses, interpretations, and more, must be taken into account prior to assessing an evidence of validity or reliability argument. It is fine to have good internal consistency as measured by the α (alpha) coefficient, but this does not provide a clear picture of the overall internal consistency argument of an assessment. For this, there are many more statistical measures to be taken into account. The remainder of this report addresses the various published areas of validity and reliability evidence from the standpoint of the American Psychological Association (APA), see [28].

2 Evidence of Validity

The APA Handbook on Testing and Assessment in Psychology, [28] has a chapter specifically devoted to measuring validity. This collection presents five sources of validity evidence, referencing the Standards for Educational and Psychological Testing, 3rd Ed., 1999, [1]. These sources are a) test content, b) relations to other variables, c) internal structure, d) response processes, and e) consequences of testing. We give a brief description of each source of evidence along with an example of a possible way of collecting such evidence.

2.1 Test Content Based Validity Evidence

In order to discuss providing evidence of validity in any form, in any setting, one must first define what is being measured. In the world of psychometric assessments, this is often referred to as defining the construct or, in the current language of the APA, defining the content domain of definition. Sireci, in [39], categorizes test content into four main areas: (1) construct definition, (2) construct relevance, (3) construct representation, and (4) appropriateness of test construction procedures.

2.1.1 Construct Definition

The construct definition is usually addressed during the early stages of development of an assessment and is usually evaluated by a group of content experts. Additionally, the content experts are usually external to and independent of the assessment development process. The content evaluators typically focus on whether important areas have been omitted or whether superfluous areas are included ([28], pp. 66).

2.1.2 Construct Relevance and Construct Representation

Construct relevance and representation are closely related and are presented together even though they represent two distinct areas of content validity evidence. Construct relevance focuses on whether each individual item is actually relevant to the construct of interest. Construct representation is a measure of whether the individual items span the full range of the content domain and avoid items that may be irrelevant to the intended construct.

It should be noted that [28] states the following on pp. 66:

> Traditional content validity studies use subject matter experts to evaluate test items for relevance and representativeness. Items that are not judged relevant to the domain are eliminated and new items are added if the experts conclude that aspects of the domain are underrepresented. ... An evaluation of the appropriateness of the construct definition focuses on how well that definition captures the *consensus understanding* of the construct.

The emphasis on consensus understanding is placed by the authors of this piece, not the authors nor editors of [28].

The main point of Sections 2.1.1 and 2.1.2 is that standards usually require an external review of these portions of the test development process by subject matter experts. As an example of the use of external reviews in this context, TTI Success Insights has piloted such a validation process by approaching subject-matter experts in industrial and organizational psychology to assess the content validity and the adequacy of the construct definitions of our behavioral assessment instrument.

The study involved having subject matter experts examine the item alignment of the behavioral assessment to determine whether descriptors align with the four domains. Armed with working definitions of each domain (DISC), reviewers were asked to categorize each of the assessment items into one of the four domains. Phase two of the review process examined the negative-positive polarity of each of the item descriptors to determine if we could identify an aversion toward a particular word even though a responder may possess a given trait. Conversely, we explored items to see if participants might select a descriptor simply because it is viewed as very positive although

it is not a trait they possess. As a result of assessing the connotative polarity of each of the item descriptors and cross walking this information with our item analysis, we are able to mitigate the effect of items that are at extreme polar ends.

2.1.3 Appropriateness of Test Construction Procedures

Appropriateness of construction procedures is evaluated based on the processes in use by the assessment developer. Items considered include, but are not limited to, adequacy of item development and selection processes, scoring procedures, and quality control. The TTI SI assessment construction template is outlined in Figure 1 presented in Appendix A.

2.2 Relations to Other Variables Based Validity Evidence

Earlier versions of Relations to Other Variables Based Validity Evidence were referred to as criterion validity, concurrent or predictive validity, or both. Technically speaking, criterion validity is a superset of a combination of concurrent and predictive validity. The section on Relations to Other Variables Based Validity Evidence in [39] tells a cautionary tale regarding collection of criterion related variables. We present two quotes from this section to highlight the level of caution presented. First starting on pp. 67,

> Cronbach (1988) and Kane (2006) distinguished between the strong form of construct validation, in which selection of validation criteria is based on construct theory, and the weak form of construct validation, in which "any correlation of the test score with another variable is welcome" (Cronbach, 1988, p.13).

The reference for Cronbach 1988 may be found in [17] and the reference for Kane (2006) may be found in [32]. The caution mentioned above refers to the above quote's clear reference to assessment developers, administrators, and users who base their validity arguments on weak levels of correlation between their constructs and *any* other variables. This approach is an excellent example of the adage that correlation does not imply causation. In both concurrent and predictive validity arguments, we require causation, not merely correlation.

Later on pp. 67 of [39], the authors state

> 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).

It is challenging to present truly predictive models based on explanatory variables in many settings and it is just as challenging when focusing explicitly on explanatory variables from a psychometric assessment. Without explicit and trustworthy metrics to measure performance, predictive validity becomes a very specialized portion of the criterion validity evidence argument.

Concurrent validity is slightly less restrictive in that we do not require as much performance data and can address, at least in part, relationships between the TTI Success Insights constructs and external variables collected through a demographics system implemented approximately one year prior to this article being authored. During the few months previous to this article, TTI Success Insights has been reviewing our data base and the demographics collection efforts to determine whether enough data is available for review of some aspects of concurrent validity between our assessment and certain demographic categories.

While the full amount of data available for consideration is more than enough, limitations are naturally imposed by the categorical nature of demographic collections such as what job one has. An example of the type of analysis planned for use in these type studies may be found in [26].

2.3 Internal Structure Based Validity Evidence

According to [39], the term internal structure refers to the dimensionality or underlying factor structure of an assessment, see pp. 72. It goes on to note that there are three main ways to measure validity evidence based on internal structure: (1) internal consistency, (2) dimensionality (factor analysis), and (3) measurement invariance.

Later in the same section of [39] the authors note that there are many different approaches to assessing dimensionality or internal structure validity. These include exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling models (MDS), and item response theory (IRT). Additionally, in [34], it is stated that

> Item analysis provides useful information related to internal structure evidence of validity, which concerns the extent to which the components (e.g., items) of the assessment are related to one another in a manner that is consistent with the intended target trait structure (Loevinger, 1957).

This quote is originally from Loevinger from her work found in [33].

The TTI Success Insights continuous improvement approach either currently incorporates or has future plans to incorporate all of the aforementioned approaches. For example, many of our assessments are undergoing reviews in which items may be updated. This approach incorporates both an exploratory factor analytic and item analysis approaches to study the individual items and their interaction with each other and how consistent they are with the intended construct being measured. Additionally, item response theory models have been built for several of our assessments and are providing an updated scoring rubric in areas in which IRT is currently applicable. A more complete discussion of the TTI Success Insights approach to IRT may be found in [10].

Our assessment review process also currently includes a confirmatory factor analytic approach as a first step to reviewing the variance explained by a construct. This is only a first step and further study of a scale then incorporates the exploratory factor analytic approach. It is the intention of the TTI Success Insights research program to incorporate fully confirmatory factor analytic models at an appropriate time in the future. More detailed information on these processes may be found in [11].

Also mentioned above is the concept of internal consistency. The authors of this paper point out that internal consistency is not a measure of validity, per se. Rather, internal consistency is a necessary, but not sufficient, condition for validity. In other words, if one could establish high levels of statistics to measure the evidence of validity, but could not first establish evidence of reliability (internal consistency), then the validity argument would fall short of its intended target. Later in this document, a discussion of TTI Success Insight's reliability study procedures are outline, see Section 4.

One approach to using internal consistency to aid in the internal structure validity argument is to consider the "alpha without" analysis. This approach takes a look at a measure of internal consistency of a scale when an item is removed from that scale. This approach helps determine

whether an item is consistent with the remainder of the scale and is a useful approach to helping determine dimensionality.

The final topic of this section is multidimensional scaling. One version of multidimensional scaling that TTI Success Insights has used over the years is our Adverse Impact study in the U.S. TTI Population. In future versions of our Adverse Impact study it is the intention of the TTI Success Insights research team to add analytic procedures suggested in [39] discussed on pp. 75-76 on evaluating invariance of internal structure via differential item functioning.

2.4 Response Process Based Validity Evidence

Response Process related evidence of validity is an interesting and in some sense exciting area of the validation argument to discuss. [39] begins its discussion of response process validity with the following, pp. 76:

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

.
As was presented in an earlier edition of this newsletter, this may no longer be the case. In an article published in October, 2018, see [12],the Target Training International research team presented the protocols for a study cross-walking between the statistical and psychometric based evidence provided by an item response theory study and evidence gathered on EEG while individuals responded to assessment items. The results of our first full study are presently being prepared for publication and will be complete sometime in late 2019. Due to the cutting-edge nature of the implication of this new protocol regarding response processing, a summary of our preliminary findings is offered at this time.

Our present response processing protocols have emerged as a result of research into decision-making pathways and is a compilation of previous brain imaging studies. To better understand this thought leadership evolution, readers are encouraged to explore the following references, including two patents that directly address this present analysis, see [3, 4, 5, 6, 7, 8, 14, 15].

By exposing prefrontal gamma asymmetry while participants are taking our assessments, we have been able to document several categories of responses to items by comparing assessment answers to real-time brain imaging. The preliminary findings include:

1. Confirmation between survey response and neurological processing
2. Items that may have socially acceptable or "correct" answers that therefore fail to match brain processing imagery
3. Mixed brain response to confusing and reverse or double negative assessment items
4. Reduced activation for a set of assessment items that many times also exhibit low item discrimination

5. Items that address the intended construct but are writing with a context that the participant does not associate

This unique approach of capturing brain activity while participants respond to assessment items is providing new insights and offers explanations to our well established psychometric analyses in was previously not possible. The interested reader may consult Appendix B for some examples from the aforementioned pilot study.

2.5 Consequences of Testing Based Validity Evidence

Conceptually, consequences of testing based validity evidence is fairly straightforward. Simply put, one needs to assess the intended and unintended consequences (positive and negative) of an assessment program. Providing an exhaustive list of all intended and unintended consequences is well beyond the scope of this document. We do provide a condensed list of some items intended to mitigate any negative consequences that may arise from the use of the TTI Success Insights suite of assessments.

In an effort to provide due diligence and set the stage for our assessments to be used as intended, we deliver learning opportunities for both new affiliates and updates for existing distributors. All new distributors (Value Added Associates or VAAs) are directed to our online Learning Management System (LMS) to expedite the learning and certification process for TTI Success Insights sciences and products. The curriculum includes five hours of video content and numerous checks for understanding that address entry level knowledge of our behaviors and motivators assessments.

This step is focused on creating the foundational knowledge-base to build upon and ensure that users understand the intended audience represented by working adults and are attentive to both the advantages and limitations of each assessment tool. Following completion of this training, VAAs are invited to attend an onsite, three day training designed to focus on the application and interpretation of our assessment scores. After completing these trainings, VAAs may register to take our two-year certification exam that assesses both content knowledge and appropriate application of our assessment scores.

Once distributors are in network, they are offered a large array of continuous learning opportunities. The list includes advanced training in all TTI Success Insights sciences, webinars, workshops, extensive debriefing guides, newsletter updates, and an annual international conference that provides participants with the opportunity to learn form each other and receive the latest product updates.

The list of all possible intended and unintended consequences related to the TTI Success Insights family of assessments is far too long to present in this format. The information in this section is intended to provide the reader with an understanding of the process TTI Success Insights has in place to address such issues.

3 Closing Remarks on Evidence of Validity

It should be noted that throughout the above remarks, care has been take to use the term evidence of validity, as well as not referring to an assessment as being valid or not. The Handbook as well as The Standards are quite clear on these topics. First and foremost, it is the *use* of the assessment and *interpretation* of its scores that are the subject of validation efforts, not the assessment itself. Second, there are many areas that evidence of validity may be gathered in. The Handbook condenses these to five main areas that have been discussed above.

Each of the five areas has a specific goal in mind. Each addresses areas of the assessment process from the development of the assessment to justification of its uses and interpretations to the cognitive functions of the individual respondent. Validation of the uses and interpretations of an assessment is a long process and the ultimate answer to the question of validity is truly the cumulative effect of many years of effort studying and understanding the whole.

4 Evidence of Reliability

The main idea behind the concept of reliability is to have a measurement that is as free from random error as possible. The key here is random error. In all measurements there is expected to be error. The concern is when one is unable to account for the error. For example, differences in scores that can be attributed to differences in ability level should not be considered errors in the sense of randomly generated errors.

Perhaps a more apt term is to consider consistency rather than the term reliable. It is the aim of all assessments to consistently measure the construct of interest. Thus, as noted in [27], consistency could be considered to be across individuals on the assessment items (internal consistency), across time for individuals (temporal or test-retest consistency), across the construct regardless of the individual items (alternative forms consistency), or across a host of other measurable facets (generalizability).

4.1 Internal Consistency Reliability

Internal consistency reliability is a form of measuring reliability or consistency that relies on a content sampling process. There is no shortage of published literature on the topic of internal consistency reliability. A handful of examples are presented for the reader, not all of which hold a view of methodologies for computing internal consistency. The following list is by no means exhaustive, [2, 16, 21, 22, 27, 30, 35, 37, 40, 41, 43]. Late in his life, Cronbach offered a critique of his own work, see [18], where he notes that the α coefficient had been given much more credit than it was due, and that better ways of measuring internal consistency existed.

One of the disadvantages of such methods is that many of these measures are essentially nothing more than scaled versions of inter-item correlation measures. Another disadvantage is that one may improve the internal consistency measure by simply adding more items of the same type already present in the assessment. Finally, there is the existence of the attenuation paradox due to Loevinger, see [33]. The attenuation paradox states that validity is not a monotonic function of reliability.

To give a relatively gentle introduction into what it means for validity to **not** be a monotonic function of reliability, simply increasing the number of items measuring the exact same portion of the construct will increase the measure of internal consistency, by definition. However, this process will also cause the scale to become highly unidimensional. In other words, one may increase the internal consistency of a scale by adding more and more of the same or similar items, the cost of which is increasing the correlation to the point of showing we are measuring but a single aspect of the underlying construct, thereby lowering the overall validity measurement.

TTI SI periodically conducts reliability studies on all of our assessments. The most recently available full reliability study for the TTI SI Style Insights assessment may be found in [24]. A view of the reliability of the TTI SI Motivation Insights assessment may be found in [23]. The interested

party may find a combination of the previous two documents in [25]. These studies are available upon request.

Further reliability studies on the TTI SI EQ assessment may be found in [11]. The Style Insights and Motivation Insights reliability studies are primarily focused on the α coefficient. In the TTI SI EQ study, the α coefficient is presented with the addition of confidence intervals. A view of internal consistency from a structural equation model perspective is also presented via McDonald's ω coefficient, with confidence intervals.

Future studies are planned to implement further and deeper analysis using the so-called "alpha (α) without" analysis. This approach has the advantage of considering the value of the α coefficient of a scale if an item is removed. This approach provides insight into the internal consistency and the internal structure validity mentioned in Section 2.3 above.

4.2 Temporal Consistency Reliability

The Handbook, see [28], refers to temporal consistency or test-retest reliability as *stability*. The key idea at play in stability of an assessment is not that any given individual taking the same assessment on multiple occasions necessarily has identical scores. Rather, it is more appropriate to consider a larger group of individuals and be more concerned with the *stability of scores* which The Handbook defines as "persistent relative standing within a group" rather than consistent absolute value of a score.

For a more complete discussion on the topic of the test-retest approach to temporal consistency, consult [27]. [11] presents a recent temporal consistency study applied to the TTI SI EQ assessment. Similarly, one may find a temporal consistency study on the TTI SI Style Insights assessment in [9].

4.3 Alternative Forms Consistency

Alternate forms consistency is just what it sounds like. In a perfect world, measuring a construct would come in many forms. As an example, one may attempt to develop two completely independent versions of an assessment designed to measure a given construct or characteristic. These two forms may then be used in separate administrations of the assessments utilizing the same group of respondents.

Ideally, the respondents would score the same in the sense of stability as defined in the previous section on temporal consistency. In the opinions of the authors of this piece, there are many issues to discuss that are most likely better left to a future article. However, we mention at least one here. The underlying assumption is that at least one of the assessments is already reliable and validated. If one of them is not already evaluated as such, what then is the point of the exercise?

Having said that, if one is working on a new generation assessment measuring a similar construct to that which has already been measured by a tool whose use and interpretation have been judged to be valid, one may consider using the older version as a measuring stick for a follow on version. All in all, the use of alternative forms of consistency appears to have many logistical and cost challenges in addition to the usual challenges of collecting appropriate data to study such concepts.

4.4 Generalizability Theory Based Reliability

Generalizability theory (G-theory) is a topic on which entire text books have been written. We cannot hope to do it all justice here. The topic is named generalizability theory because it is intended to be a generalization of the concept of internal consistency reliability in the sense of Cronbach. In fact, Cronbach was one of the driving forces behind the development of the theory in the 1970s, see [19]. Some excellent resources may also be found in any of the following [13, 38, 42], to name a few.

In addition to identifying the areas where variance in responses arises through the G-theory, there are two important metrics associated with this approach. The first, called the generalizability coefficient, is analogous to the classical test theory measure of reliability discussed in Section 4.1. According to [42], this generalizability coefficient focuses on relative decision concerns such as identification of the top $x-\%$ of a group. The second metric is the so-called dependability coefficient and is related to absolute decisions, such as a pass-fail criteria, that focus on an individual's performance independent of others' performances. According to [13], classical test theory has no analogous metric to the dependability coefficient from G-theory.

TTI Success Insights applies G-theory in situations that warrant such studies. An example of this application may be found in Chapter 9 of [11]. This chapter represents multiple applications of G-theory across multiple time periods using both one and two-facet G-theory models to explore the relationships between assessment respondents and their scores relative to time, and between respondents, time, and individual items. The topic of study is the TTI SI EQ assessment. Generalizability and dependability coefficients are presented in addition to tables with Generalizability analysis of data information.

4.5 Closing Remarks on Evidence of Reliability

Evidence of consistency or reliability is an important part of evaluating the overall effectiveness of the use and interpretation of assessment scores. Similar to discussing validity in terms of the use and interpretation of scores based on an assessment, we also speak of reliability in these terms. In other words, one should discuss the concept of evidence of reliability of the scores of an assessment or the interpretation or use of those scores and not necessarily the assessment itself.

Closing Remarks

Evidence of reliability and validity go hand in hand in establishing the argument for validity of the use and interpretation of an assessment. Such an argument cannot be created solely based on one aspect of either area. Establishment of these arguments takes time and solid performance of the results of the assessment over many years.

One misconception that appears to have been propagated through the commercial assessment community is that a measure of internal consistency such as the so-called Cronbach's alpha is a proxy for a validation argument. This is a false argument. Said in one way, reliability is a necessary but not sufficient condition for validity. Said in another, reliability is an important aspect of establishing the validation argument, but is by no means the entire argument.

There are many resources world wide that deal with the general area of assessments and their use and interpretation, and specifically in the area of psychological testing or psychometrics. As a result of our global positioning as an assessment company, we are aware of and examining additional

insights that may be gleaned from other nations and larger international bodies that have their own assessment guidelines. At present, this list of resources includes but is not limited to:

- The European Federation of Psychologists' Associations (EFPA)
- The International Test Commission Assessment Guidelines on Test Use
- The Australian Psychological Society (APS) Supplement to Guidelines on the Use of Psychological Tests
- The British Psychological Society (BPS) Level A and Level B Standard for Occupational Test Use
- The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Standards for Educational and Psychological Testing
- The American Association for Counseling and Development (AACD) Responsibilities of Users of Standardized Tests
- The Canadian Psychological Association (CPA) Guidelines for Educational and Psychological Testing

References

[1] 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.

[2] Douglas G. Bonett and Thomas A. Wright. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, 2014.

[3] Bill J. Bonnstetter, Ronald J. Bonnstetter, Dustin Hebets, and Tom F. Collura. Validation process for ipsative assessments, 2015. US Patent 9,060,702.

[4] Bill J. Bonnstetter, Ronald J. Bonnstetter, Dustin Hebets, and Tom F. Collura. need a title, 2016. Canadian Patent 2,808,691.

[5] R. J. Bonnstetter, D. Hebets, and N. L. Wigton. Frontal gamma asymmetry in response to soft skills stimuli: A pilot study. *NeuroRegulation*, 2(2):70–85, 2015. http://dx.doi.org/10.15540/nr.2.2.70.

[6] Ronald J. Bonnstetter and Tom F. Collura. Brain Activation Imaging in Emotional Decision Making and Mental Health: A Review Part 1. *Clinical EEG and Neuroscience*, 2019. In review.

[7] Ronald J. Bonnstetter and Tom F. Collura. Brain Activation Imaging in Emotional Decision Making and Mental Health: A Review Part 2. *Clinical EEG and Neuroscience*, 2019. In review.

[8] Ronald J. Bonnstetter, Tom F. Collura, Dustin Hebets, and Bill J. Bonnstetter. Uncovering the belief behind the action. *NeuroConnections*, Winter 2013. Newsletter of the International Society of Neurofeedback and Research and Applied Psychophysiology and Biofeedback Neurofeedback Division. McLean, VA.

[9] Ronald J. Bonnstetter and Eric T. Gehrig. TTI Success Insights Style Insights 2016 Temporal Consistency Report. Technical report, TTI Success Insights, 2016.

[10] Ronald J. Bonnstetter and Eric T. Gehrig. A Graded Response IRT Model for Analysis of the TTI Success Insights EQ Assessment. Technical report, TTI Success Insights, 2019.

[11] Ronald J. Bonnstetter and Eric T. Gehrig. TTI Success Insights Emotional Quotient 2019 Technical Manual Version 1.0. Technical report, TTI Success Insights, 2019.

[12] Ronald J. Bonnstetter, Eric T. Gehrig, and Dustin Hebets. Response Process Validation Protocol using Neurophenomenological Gamma Asymmetry. *NeuroRegulation*, 2018. Accepted.

[13] Robert L. Brennan. *Generalizability Theory*. Statistics for Social Science and Public Policy. Springer, New York, 2001.

[14] 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.

[15] 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.

[16] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

[17] Lee J. Cronbach. Five Perspectives on the Validity Argument. In H. Wainer and H. I. Braun, editors, *Test Validity*, pages 3–17. Hillsdale, NJ: Erlbaum, 1988.

[18] 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.

[19] 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.

[20] Robert F. DeVellis. *Scale Development: Theory and Applications*. Sage, Thousand Oaks, CA, 4th edition, 2017.

[21] 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.

[22] 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.

[23] Eric T. Gehrig. 2017 Reliability Study: TTI Motivation Insights. Technical report, TTI Success Insights, 2017.

[24] Eric T. Gehrig. 2017 Reliability Study: TTI Style Insights. Technical report, TTI Success Insights, 2017.

[25] Eric T. Gehrig. 2017 Reliability Study: TTI Talent Insights. Technical report, TTI Success Insights, 2017.

[26] Eric T. Gehrig. Classification of Serial Entrepreneurs via Logistic Regression: A Case Study. White Paper, October 2017.

[27] 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.

[28] 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.

[29] Richard L. Gorsuch. *Factor Analysis*. Lawrence Erlbaum Associates, 2nd edition, 1983.

[30] Louis Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 42(4):252–282, 1945.

[31] Robin K. Henson, Roberto M. Capraro, and Mary Margaret Capraro. Reporting practice and use of exploratory factor analysis in educational research journals. In *Annual Meeting of the Mid-South Educational Research Association*. Mid-South Educational Research Association, November 2001.

[32] M. Kane. Validation. In R. L. Brennan, editor, *Educational Measurement, 4th Edition*, pages 17–64. Washington, D.C.: Rowman & Littlefield, 2006.

[33] Jane Loevinger. The attenuation paradox in test theory. *Psychological Bulletin*, 51:493–504, 1954.

[34] 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.

[35] Gjalt-Jorn Y. Peters. The alpha and omega of scale reliability and validity. *The European Health Psychologist*, 16(2):56–69, April 2016.

[36] Keenan A. Pituch and James P. Stevens. *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. Routledge, New York and Abingdon, 6 edition, 2016.

[37] William Revelle and Richard E. Zinbarg. Coefficients alpha, beta, omega, and the glb: comments on sijtsma. *Psychometrika*, 74(1):145–154, 2009.

[38] Richard J. Shavelson and Noreen M. Webb. *Generalizability Theory: A Primer*. Sage, Thousand Oaks, CA, 1991.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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 TTI Success Insights Assessment Construction Template

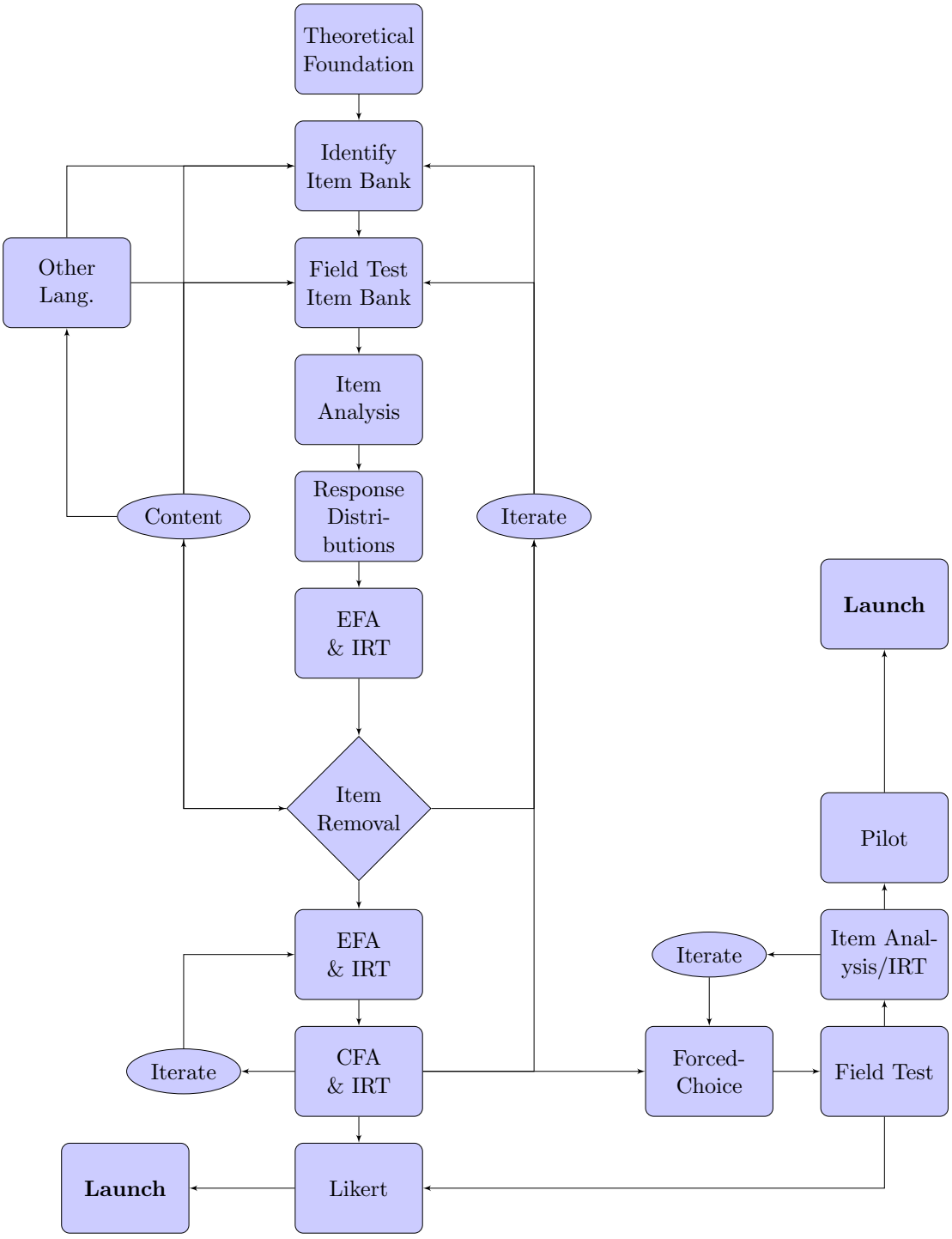

Figure 1: Assessment Development Flow Chart

B TTI Success Insights Response Process Pilot Study Examples

This appendix provides several examples comparing brain imaging processes with assessment responses.

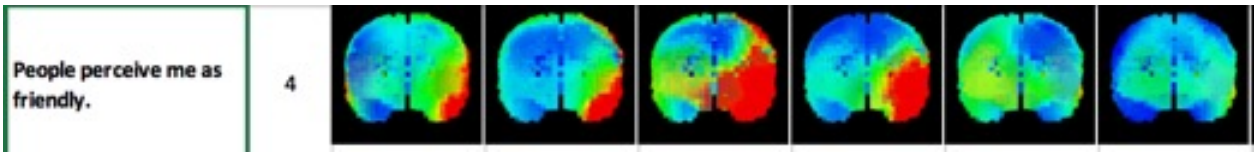

Figure 2: Confirmation Between Survey Responses & Brain Imaging

The scale response of 4 (using a Likert-type scoring with 1 representing strong disagreement and 6 representing strong agreement) and the left prefrontal gamma approach response show agreement between the response provided and the brain imaging. Each images is a linear progression at $\frac{1}{8}$ second intervals.

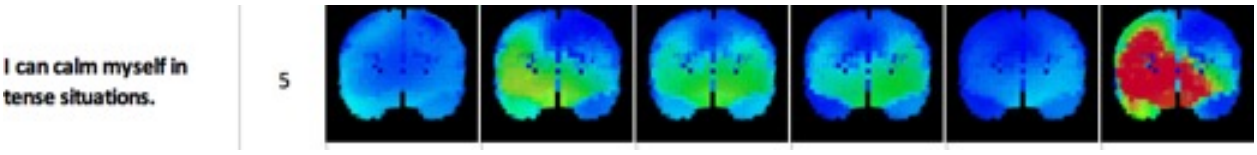

Figure 3: Possible Socially Acceptable Response Brain Activity

While the participant responded with 5, a fairly strong agreement with the item, the brain shows a strong avoidance response within $\frac{3}{4}$ second after exposure to the statement.

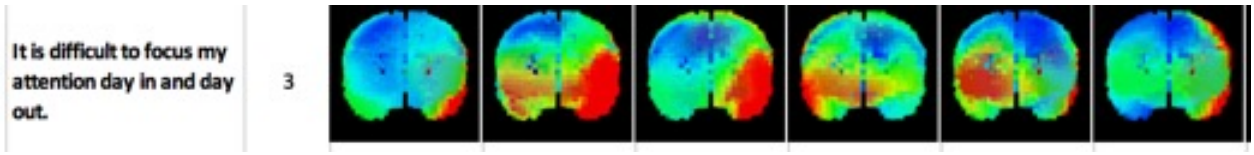

Figure 4: Brain Response to Confusing Items

Both the neutral response on the survey (3 representing a slightly negative view of the item) and the fluctuation in the brain response suggest confusion, possibly with the portion of the item referring to day in and day out.

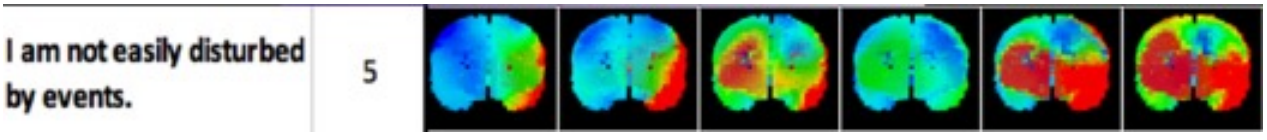

Figure 5: Brain Response to Negative Items

A pattern of fluctuation has emerged when participants are confronted with negatively stated items. While the survey response indicates agreement, the brain starts with agreement but moves to a

strong mixed response of both acceptance and avoidance.

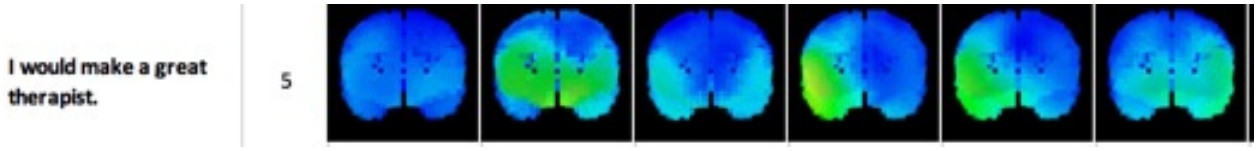

Figure 6: Item Context With Which Participant Does Not Associate

When confronted with an unexpected context (being a therapist) the participant's response lack emotional intensity. This reaction is helping identify items that may lack directionality.

C TTI Success Insights Internal Consistency Examples

Table 1: Common Reliability
Coefficient Interpretation, see [20]

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

Table 2: Sample α Coefficient Data
TTI SI Style Insights

Lang./Scale	D	I	S	C
English AU/NZ	0.89	0.86	0.85	0.84
English UK	0.89	0.85	0.85	0.84
English US	0.89	0.86	0.85	0.84
German	0.91	0.85	0.87	0.86
Spanish Americas	0.86	0.84	0.80	0.81

Table 3: Sample α Coefficient Data
TTI SI Motivation Insights

Lang./Scale	The	Uti	Aes	Soc	Ind	Tra
English AU/NZ	0.85	0.82	0.81	0.88	0.84	0.82
English UK	0.84	0.78	0.78	0.86	0.84	0.80
English US	0.85	0.82	0.82	0.88	0.84	0.83
German	0.81	0.72	0.85	0.84	0.87	0.70
Spanish Americas	0.80	0.81	0.67	0.87	0.77	0.76

D TTI Success Insights G-theory Model Templates

Table 4: Scale X Test-Retest as a One Facet G-Theory Model

Individual	**Items**				
Respondent	**1**	**2**	$\cdots$	**m-1**	**m**
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,(m-1)}$	$X_{1,m}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,(m-1)}$	$X_{2,m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$X_{n,1}$	$X_{n,2}$	$\cdots$	$X_{n,(m-1)}$	$X_{n,m}$

Table 5: 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	$X_{n,1,1}$	$X_{n,2,1}$	$\cdots$	$X_{n,m,1}$	$X_{n,1,2}$	$X_{n,2,2}$	$\cdots$	$X_{n,m,2}$

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

Figure 7: Two Facet Defining Equations for $P \times T \times R$ Model