Some Thoughts on Current Consensus Views on Evidence of Reliability and Validity in the Psychometric Assessment World

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

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

August 30, 2019

Abstract

There are many challenges presented when working in the setting of the private sector while largely applying academic concepts. There can be, at times, widely varying opinions of what is acceptable and what is not in a given space. These widely varying opinions often come from the academic community, as well as from the business community, with little in the way of solutions offered. This manuscript attempts to highlight some of these areas in the assessment world. In many cases, organization such as the American Psychological Association, provide general guidelines without providing specific guidance. As an example, we discuss the concepts of acceptable levels of reliability coefficients, as measured by the so-called "alpha" coefficient where after several decades there remains a lack of consensus. Similarly, we discuss the topic of variance explained and the lack of any relevant guidance other than that fact that it should be reported. This manuscript also touches on the areas of correlation versus causation, self-report assessments, and forced-rank versus Likert style response formats. The intention of this manuscript is to highlight areas in which the assessment community, in the opinion of these authors, would greatly benefit from more specific guidance than the usual dependent on context statements that populate the literature on the topic.

1 Introduction

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 [16], [17], and [18]. 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 [2]. 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.

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

Among all the guidance provided by these noted organizations and academic literature 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 [31], 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 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.

2 Thoughts on Evidence of Reliability

We go more in-depth into the history of reliability coefficients in the section on internal consistency in [7]. 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 [23]. 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 [13] and [19]. [13] is a bit more useful in that it discusses several different approaches and offers some suggestions as to directions one could take. [19] is a nice exposition on an empirical analysis of the use, or misuse and misunderstanding, of the α coefficient specifically.

[13] 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, [19] 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 [13] and [19], 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 [13] and [19] present slightly differing views, they generally agree that [23] first proposed a "reliability formula (called KR-20) that can be used for data on dichotomously scored items...", see [13]. [19] states that "Alpha was introduced by Kuder and Richardson (1937) for dichotomously scored items". We point out this difference given that [13] does not refer the result of [23] as "Alpha" while [19] 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.

[13] traces the history of reliability coefficients to Spearman and Brown, both independently publishing in the same issue of the *British Journal of Psychology*, see [32] and [12]. The formula published has since become known as the Spearman-Brown formula. [19] begins their discussion with [23]. 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 [13] is introducing.

Table 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 [13] was published in 2015 and [19] 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 1, above, is presented in [14] and reproduced here for purposes of the current discussion. Additionally we present the following quote from [5]:

> 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 [27] stating that 0.70 is acceptable during development, 0.80 for "basic research scales", and 0.90 for clinical settings are appropriate minimums. [29] recommends a minimum of 0.80 without the same restrictions as [27]. As a final example, [22] suggests values of 0.95 for clinical settings. [5] 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 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 of the use of 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 [16] 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 [31] 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.

[2] 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 [2] 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 [2], 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 the 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 [31].

> 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 manuscript in Section 1 and the 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 2019 or early 2020. We have already had the protocol paper accepted and published, see [8].

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 [31] 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 [31] 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, [7]. 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 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 [7]. 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 [31]. 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 Thrustonian IRT model to help in the analysis of both the response processing and the internal structure evidence of validity. We do note that, for diagnostic purposes, we have employed exploratory factor analytic approaches on the Style Insights scales individually. While this approach is not directly applicable to establishing internal structure validity, it has been invaluable in identifying those items that perform well and those that do not. What it cannot do is give us a reason for why certain items underperform given that they are also influenced by the existence of three other items in the forced-rank frame. See Section 6 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 [2], 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 mul-

> tivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the score variability attributable to any other identified dimension, or showing that a single factor adequately accounts for the covariation among test items.

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

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

We briefly discussed the evidence of validity based on relations to other variables in the introduction to this section. In Section 4, 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 [31] 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 [31], they do take the stance that

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

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

4 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 1. The correlation between the two variables under consideration is just under 95% (94.71% 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.

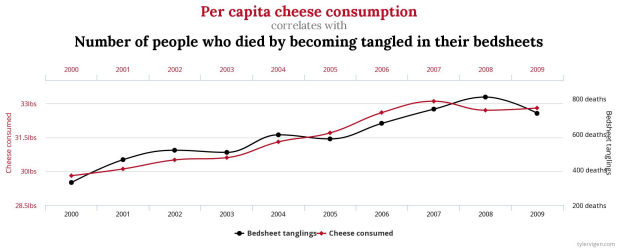

Figure 1: Bad Correlation v Causation Example

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 that $|0.40|$ 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.

5 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 [30]:

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. [6], [24], and [28] 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 [6], 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 [3], [4], [15], and [33] 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.

6 Forced-rank Assessment v. Likert 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 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 and presented in more detail in section on Thurstonian IRT models in [7].

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).

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 as-

sessment 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 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 I 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 [21], 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 [21] 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 [21] 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 [21]. 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 [21]. The work presented in [9, 10, 11, 25, 26], 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 the business community and it is appreciated. The not as good news is that we are still left to interpret the output.

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.

7 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 later, leaving the former to another time.

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 face with everything from assessment 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 place on such examinations based on the simple fact that one must differentiate themselves if they wish to find the best opportunities.

We state that last comment in light of the fact that secondary education grades are highly inflated, see [20]. 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 the average secondary school, a student averages 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 measure, 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 the 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, there was a 0.6% difference in graduation rates.

The fact that we have cited only two references aside, the issue is a real one. We have a system in which assessments (either self-report, supervisory, or academic) are required for us to advance. 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. We again pose the question from the introduction. Is it possible to validate the use of any assessment in a system that appears to flaunt the existence of such assessments?

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] 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.

[3] 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.

[4] 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.

[5] Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. The Guilford Press, New York, 2018.

[6] 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.

[7] 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.

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

[9] Anna Brown and Albert Maydeu-Olivares. Item response modeling of forced-choice

questionnaires. *Educational and Psychological Measurements*, 71(3):460–502, 2011.

[10] Anna Brown and Albert Maydeu-Olivares. Fitting a Thurstonian IRT model to forced-choice data. *Behavioral Research Methods*, 44:1135–1147, 2012.

[11] Anna Brown and Albert Maydeu-Olivares. How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18:36–52, 2013.

[12] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910.

[13] Eunseong Cho and Seonghoon Kim. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2):207–230, April 2015.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Scott Jaschik. High school grades: Higher and higher. https://www.insidehighered.com/admissions/article/2017/07/17/study-finds-notable-increase-grades-high-schools 2017. Inside Higher Ed.

[21] 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.

[22] Robert M. Kaplan and Dennis P. Saccuzzo. *Psychological Testing: Principles, Applications, and Issues*. Wadsworth, Belmont, CA, 2001.

[23] G. F. Kuder and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937.

[24] 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.

[25] Albert Maydeu-Olivares and Ulf Böckeenholt. Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, 10(3):285–304, 2009.

[26] Albert Maydeu-Olivares and Anna Brown. Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, 45:935–974, 2010.

[27] Jum C. Nunnally. *Psychometric Theory*. McGraw-Hill, New York, 2nd edition, 1978.

[28] 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.

[29] Tenko Raykov and George A. Marcoulides. *Introduction to Psychometric Theory*. Routledge, New York, 2011.

[30] 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.

[31] 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.

[32] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910.

[33] 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.