2016

Detecting Insufficient Effort Responding: An Item Response Theory Approach

Tyler Douglas Barnes
Wright State University

DETECTING INSUFFICIENT EFFORT RESPONDING: AN ITEM RESPONSE THEORY APPROACH

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science

By

TYLER DOUGLAS BARNES
B.S., Eastern Kentucky University, 2013

2016
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

05/23/2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Tyler Douglas Barnes ENTITLED DETECTING IER: AN ITEM RESPONSE THEORY APPROACH BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

David LaHuis, Ph. D.
Thesis Director

Scott Wattaminuk, Ph. D.
Graduate Program Director

Debra Steele-Johnson, Ph. D.
Chair, Department of Psychology

Committee on Final Examination

Gary Burns, Ph. D.

Nathan Bowling, Ph. D.

Robert E. W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

Abstract

Barnes, Tyler Douglas. M.S. Department of Psychology, Wright State University, 2016.
Detecting IER: An Item Response Theory Approach

Insufficient Effort Responding (IER) is prevalent enough in self-report data to cause issues with construct validity. There are many ways to detect IER, but they are less than ideal as they each detect different forms of IER. I compared an Item Response Theory (IRT) approach consisting of the l_z person-fit statistic and the Person Fluctuation Parameter (PFP) to longstring, non-consecutive longstring, even-odd split, and psychological synonyms indices. I simulated 3200 samples with one of four types of random responding: consecutive responding, non-consecutive patterned responding, random responding following a normal distribution, and random responding following a uniform distribution. Also, I generated an additional sample that consisted of all types of IER examined within this study. I found that the IRT methods are able to detect IER considerably better than the other indices, excluding using the longstring method to detect consecutive responding. As such, they are robust enough to detect most forms of IER. I conclude that using IRT approaches after removing the obvious IER cases with the longstring index is the best way to detect IER.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Detecting IER: An Item Response Theory Approach

There has been a renewed emphasis on the quality of data provided by self-report responses to survey measures. The primary focus has been on identifying individuals who provide unusable data because of careless or insufficient effort responding (IER; Huang Curran, Keeney, Popski, & DeShon, 2012). Effortless responding adversely affects the factor structure (e.g., Schmitt & Stults, 1986; Woods, 2006) and criterion-related validity (McGrath, Mitchel, Kim, & Hough, 2010) of the test, thus limiting the usefulness of the measure. Therefore, it is important to identify and remove these respondents to maintain the psychometric integrity of the test.

Due to the affects that IER can have on data quality, researchers have devised some ways to detect people who engage in IER (e.g., Huang et al., 2012). However, these methods each have their flaws and are less than perfect detectors of IER. One possible reason for this is that they are only able to detect one kind of IER although there are different behavioral manifestations of IER. One possible fix for this is by using Item Response Theory (IRT) approaches to detecting IER. Theoretically, IRT approaches should be able to detect all forms of IER and thus, be strong detectors of IER.

Below, I begin by discussing the role of self-reported surveys in research and in selection procedures. Within these surveys, IER is prevalent at levels that can lead to improperly made decisions. I discuss statistical techniques that researchers use to detect IER or cases. I expand on on previous research by Huang et al. (2012) and Meade and Craig (2012) by comparing the effectiveness of the more common statistical approaches for IER detection with IRT approaches. Researchers have examined the effectiveness of each IER indicator (e.g. Huang et al. 2012; Meade & Craig, 2012), but there has yet to be a study examining IRT in conjunction with the other indices. The purpose of my study is to compare the IER indices in effectiveness and determine which index is the best indicator of IER.

The Use of Self-Report Surveys

Self-report is one of the most common ways to assess many constructs used in research and selection procedures (Spector & Brannick, 2009). This stems from the many advantages of using self-report data. First, they are cheap and easy to use. It is relatively easy to post an internet survey on various websites and research management systems. Many websites code the data into a useable form, such as an electronic spreadsheet, for data analysis. Also, researchers can distribute self-report surveys to massive numbers of participants. Peer-reviewed self-report surveys have demonstrated reliability and validity. It is rare to find a statistical package in which one cannot compute internal consistency indices such as Cronbach's Alpha or perform correlations of multiple test administrations to demonstrate test-retest reliability. As such, self-report surveys usually assess constructs on a scale that is readily useable for reliability and validation analyses. Finally, psychology researchers measure some constructs that require asking the participant. These constructs can include feelings, emotions, deviant behaviors, and other covert behaviors and constructs that researchers cannot observe. In these instances, the only option to measure the constructs is to ask the participant.

Self-report methods have many advantages, but they also have some flaws. One major flaw is that self-report measures are prone to aberrant responding (Karabatsos, 2003). Aberrant responding is defined as a response pattern on the test that does not reflect the true ability of the examinee. Typically, researchers discuss five domains of aberrant responding (creative responding, cheating, lucky guessing, careless, and random responding), but I limit our discussion to the latter two.

Random responding occurs when the test taker selects response options unsystematically (Karabatsos, 2003). Careless responding encompasses every other kind of response pattern associated with inattentiveness. Careless responding includes two components: response sets and inconsistent responding. Response sets occur when the test takers follow a pattern when they answer, but the pattern is unrelated to the test taker's underlying level of the assessed construct (e.g., on a 5-point graphic rating scale, responding with all fives, or 1,2,3,4,5,1,2,3,4,5, etc.). Inconsistent responding is when a test taker answers related questions differently, but non-randomly (e.g., a participant reports low extraversion and reports high extraversion in the same test). The behaviors of random and careless responding might differ, but the underlying mechanism that drives these two behaviors is the same: inattentiveness to the test.

Researchers have consolidated careless responding and random responding into one category, coined insufficient effort responding (IER; Huang et al., 2012) This IER definition subsumes both careless and random responding. In this study we assess IER as defined by both consistent responding (both consecutive responses and non-consecutive responses with a pattern) and inconsistent/random responding following Meade & Craig's (2012) recommendations by using both a random normal and random uniform distribution from which to draw random responses.

Prevalence of IER

IER is prevalent in self-report data, especially in online administration of surveys (Beach, 1989). Johnson (2005) reported a base rate prevalence of 3.5% of his sample as

IER cases. Meade and Craig (2012) reported that 10 to 12%. Both studies use a conservative classification of IER (participants responded to a large, however unspecified, portion of the test with IER). Researchers tend to find that most people, 50 to 73%, report some form of IER on at least one item (Baer, Ballenger, Berry, & Wetter, 1997; Berry et al., 1991).

Issues Associated With IER

IER of any prevalence is problematic in psychometric research for at least three reasons: it affects reliability, criterion-related validity, and construct-related validity of self-report measures. When test takers respond randomly, those responses can attenuate internal consistency coefficients such as Cronbach's Alpha by answering inconsistently. When someone responds in a response set, the effect on reliability is unpredictable because the coefficients can be strengthened, attenuated, or the same as the true coefficient (Meade & Craig, 2012). Regardless of the resulting reliability coefficient, it will be inaccurate. Furthermore, removing IER cases can improve internal consistency (Huang et al., 2012).

The same is true for criterion-related validity coefficients. Hough et al. (1990) demonstrated that random responding attenuated correlations between personality and job performance variables. Logically, IER can strengthen relationships between two variables. For example, if you are measuring two constructs and both measures are susceptible to IER, it is possible that the measures both capture IER. This results in stronger relationships due to common method variance.

People who engage in IER can report differently than the careful responder and therefore will be considered an outlier. This is evident from the Mahalanobis *D* outlier analysis that researchers can use to detect IER cases. Because these people are considered outliers, they can have a strong effect on regression coefficients (Stevens, 1984). Stevens (1984) argued that 1 or 2 outliers can have a strong influence on regression coefficients. This has a detrimental effect on the validity of findings and utility (Stevens, 1984).

Similar to IER's effects on criterion-related validity, it can affect other areas of construct validity. When IER is present in 10% of a sample, unidimensional scales fit a two-dimensional model better than a unidimensional model (Schmitt & Stults, 1985). As such when removing cases with IER, the unidimensionality of a scale is improved (Huang et al., 2012). Furthermore, it is unethical and invalid to interpret test results when researchers are unsure that the test measures the intended construct.

It is important to address a practical issue with IER. The goal of any selection procedure (school, workplace, etc.) is to distinguish between people who might perform well on a task and those who might perform poorly. This distinction is frequently made by using scores on tests. The use of the GRE, SAT and/or ACT for graduate and undergraduate admissions in colleges and tests of personality, job experience, and/or general mental ability for hiring are just a few examples of the prevalence of using test scores to make decisions. Because IER can lead to spuriously high or low scores, practitioners should not underestimate the importance of obtaining "true" scores on tests to improve prediction accuracy (Karabatsos, 2003).

Review of Detection Methods

Because of these issues due to IER, researchers have developed many ways to detect IER. These different techniques can be clumped into two different categories:

before (*a priori*) and after (*post hoc*) administration of the survey (Meade & Craig, 2012). The techniques applied before test administrations involve making items/scales to directly or indirectly measure IER. These *a priori* methods will not be used in this study and will not be discussed further. The *post hoc* techniques are statistical indices computed after the data are collected. Some of the popular techniques are reviewed below.

Traditional *Post Hoc* Detection Methods

Some of these methods are complex, but have the advantage of not increasing survey length. These methods can detect different types of IER behaviors that some of the a priori methods, but all have the disadvantage of not having clearly defined cut-off scores for classification into IER or non-IER cases. I will discuss only the indices that I used in the study. For a more comprehensive review, I recommend Meade and Craig (2012).

Long String. Long string measures detect people that give the same response, consecutively, for many items. Also, long string methods can detect people who follow small response sets (e.g., 1,2,1,2,1,2, etc.). There are two methods within the long string technique: maximum long string and average long string.

The maximum long string index involves finding the number of times a test taker gives a consecutive response or set of responses. The index is that number. For example, if a participant answers with a "1" ten times consecutively, the index is ten (not one). In average long string, researchers must give participants clusters of items on separate pages. The researchers compute the maximum long string from each page and then average the indices together as a measure of IER throughout the entire survey.

This method detects only a small portion of IER cases. Meade and Craig reported that 2% of 436 people followed a response pattern conducive to detection by long string methods (2012). While it is important to ensure data quality, 2% is less than the conservative base rate of IER as proposed by Johnson (2005). In summary, this method does not seem to be effective at detecting the majority of IER cases.

Consistency indices. There are three common consistency indices: psychological antonyms, psychological synonyms, and even-odd split or person reliability. The psychological antonyms index involves correlating all combinations of item pairs to tease out the item pairs that have strong negative correlations. The index is computed by correlating the observed scores between the antonym pairs for each person. A stronger negative correlation for each person indicates more consistent responding. The psychological synonyms index follows the same process as the psychological antonyms, but instead, strong positive correlations between item pairs are used for the index. The resulting correlation coefficient from the index should be strong and positive to indicate consistency. For both of these indices, the convention is to find 30 pairs of items with a correlation of at least .60 or -.60 (synonyms and antonyms, respectively) from which to draw these indices.

The even-odd split index is computed by first reverse coding negatively coded items. Then the researcher sums the even numbered items within each dimension of the test. The same is done for the odd numbered items. The even and odd summed scores within each dimension are the pairs used for the within-person correlation. Therefore, if a test has five dimensions, the within person correlation will come from a correlation with 5 pairs. A stronger positive correlation implies consistent responding.

Huang et al. found that the ability for the consistency indices depends on the cut off scores (2012). The psychological antonym scales ability to detect IER ranges from detecting 20 to 45% of overall IER cases with a 95% specificity (accurate 95% of the time) cut off. The even-odd split ranged from detecting 16 to 46% of IER cases overall. The best came from the previously established cut-off of a correlation of r =.30. "Obvious" IER cases were easier to detect than suspect cases (Huang et al., 2012).

Meade and Craig found that the psychological synonyms detected 73.3 % of IER cases and the odd-even split method identified 64.4% in empirical data (2012). They also found, in simulated data, the odd-even split out performed both the psychological synonyms and antonyms regardless of the prevalence of IER within the sample and within the survey (Meade & Craig, 2012). It is important to note that all of the consistency indices decline in detection ability when participants answer with IER on small portions of the survey.

These indices have their weaknesses. Both the antonym and synonym indices require extremely large tests to find the 30 pairs of related items with a correlation of .60 or -.60 needed for an accurate estimate of the correlations. Failure to find the required number of these pairs can render the correlation uninterpretable because a slight difference in one pair will change the index score drastically. This issue is a problem with even-odd split as well. A researcher must administer a large test with a large number of subtests in order to obtain a stable index.

Item Response Theory (IRT)

Item response theory (IRT) is a framework of mathematical models which I believe underlies how a person responds to an item. Unlike in Classical Test Theory, IRT assumes that the relationship between a person's response and their underlying trait/ability (or theta) is non-linear and that items measure different levels of theta differently.

IRT models give researchers a lot of information about the items. Figure 1 is an illustrative example of the information gained from three dichotomous items. These curves are called Item Response Curves (IRCs). These items are modeled from the two-parameter logistic model seen below:

$$Probability(x|\theta, \alpha, \beta) = \frac{exp^{(\alpha(\theta-\beta))}}{1 + exp^{(\alpha(\theta-\beta))}}$$

From theta (x-axis) and the item properties, I can calculate the probability of a correct response (y-axis). Each of the items contains two item properties, or parameters: discrimination (α) and location or difficulty (β).

Discrimination is described by how much of a change of ability induces a change in probability of a correct response. A higher discrimination parameter implies that it is easier to distinguish between people of similar ability. This is useful in selection decisions where many applicants have similar ability and the goal is to find the "best of the best". For example, assume that I gave two people a test item rated on a scale from one to 100. One person scores a 95 and the other scored a 96. If the item is not highly discriminating, then the underlying ability estimates would be comparable. If the item is highly discriminating, then the underlying ability estimates are distinct.

Discrimination is illustrated by the slope of the IRCs. High discrimination (Item 2 and Item 3 in figure 1) implies a steep slope and low discrimination (Item 1 in figure 1) implies a flatter slope. Therefore, in a selection decision, it would be more prudent to use Item 2 and/or Item 3.

Another view of the discrimination parameter is that it is not an item characteristic, but a function of how the test taker interprets the item, or how well the responses on a test reflect inconsistent interpretation of the test taker's own ability estimates (Lumsden, 1977; Ferrando, 2009). For example, on a test of conscientiousness, on one item a person may believe that he/she is high in conscientiousness and report thusly. Upon reading the next item, the person might re-evaluate his/her own level of conscientiousness and believe that his/her own level of conscientiousness is different from what he/she reported on the previous item. This leads to inconsistent responding, which is quantified by Person Fluctuation Parameter (PFP). The IRT model that incorporates this alternative view of the discrimination parameter substitutes the item discrimination parameter for the PFP.

The location, or difficulty, parameter is defined by how much theta is needed to have a 50% probability of a correct response on an item. This definition assumes a dichotomous item where there is a correct response. In Figure 1, Items 1 and 2 have the same difficulty parameter as the two curves intersect at 50% probability. Item 3 would be considered more difficult, or has a higher *б* parameter, because the amount of ability needed to achieve this 50% probability is more than the other two items.

While the above examples are described with dichotomous items, these definitions and the underlying mechanisms are similar with polytomous items. The interpretation of theta is exactly the same. The interpretations of the discrimination and location parameters are similar, but there are separate probability calculations for each response option (i.e. if you have a 5 point likert-type scale, you will calculate a probability from all five response options). You will still have one discrimination parameter, but you will have multiple location parameters; a parameter for each response option minus one. For example, if you have five response options you will have four location parameters.

In the graded response model, an IRT model used for polytomous items, the calculations for the probabilities of each response option involves two steps. The first step is calculating a Boundary Response Function (BRF) for each location parameter and each ability estimate using the following formula:

$$BRF = \frac{exp^{(\propto(\theta-\beta))}}{1 + exp^{(\propto(\theta-\beta))}}$$

This results in four BRFs for each person, assuming five response options. The first BRF is interpreted as the probability of responding with a 2, 3, 4, or 5 (assuming that you have 5 point likert-type scale). The second BRF is interpreted as the probability of answering with a 3, 4, or 5. The third BRF is interpreted as the probability of answering with a 4 or 5. This pattern continues for each BRF.

These BRFs give researchers probabilities, but the probabilities are not specific enough for practical use. For example, the third BRF tells us the probability of answering with a 4 or 5, but this is not the probability of a specific response. The second step is the calculation of the probabilities for a specific response option. Assuming a 5 point likert-type scale, the formulas are as follows:

$$\textit{Probability Responding with a 1} = 1 - BRF_1$$

$$\textit{Probability Responding with a 2} = BRF_2 - BRF_1$$

$$\textit{Probability Responding with a 3} = BRF_3 - BRF_2$$

$$\textit{Probability of Responding with a 4} = BRF_4 - BRF_3$$

$$\textit{Probability of Responding with a 5} = BRF_4$$

From the probabilities obtained, researchers can calculate the likelihood of each response pattern. Researchers can use this information to calculate a person-fit statistic.

Item Response Theory person fit statistics. A person-fit statistic is an index of how likely is a person's response pattern or how much a person deviates from what the model expects. Although these two conceptualizations are different, they converge upon answering the same question. Does this person match our expectations based on the proposed IRT model? Within the context of my study, person-misfit might be an indication of IER.

IRT has a large advantage over the other methods because the person-fit statistics should detect any kind of IER. All IER behaviors are considered deviations from prior expectations or the response model. In long string, IER is deviation from the researcher's expectations of how often a consecutive response is appropriate. In Mahalanobis D, IER is deviation from the mathematical centroid. In the consistency indices, IER is deviation from the expectation of answering similar items similarly. Because all person-fit indices compare expectations derived from an IRT model to observations, this encompasses all of the definitions posited by the other indices. This means that IRT person-fit statistics should detect all forms of IER.

There are many different person-fit statistics and they differ in their ability to detect careless responding (Meijer, 1996). Some are capable of detecting random responding and other forms of IER, and some are better at detecting people who do not understand the instructions or other forms of aberrant responding (Karabatsos, 2003). This feeds into the advantage of IRT having the ability to detect all forms of aberrant responding including IER.

The literature on the effectiveness of the person-fit statistics is contradictory. These differences partially stem from: 1) the difference in test lengths, 2) the number of misfitting responses within a single case of aberrant responding, 3) the prevalence of misfitting response patterns within the sample, and 4) the vast number of different statistics.

Longer tests tend to lead to better detection rates of person-fit statistics (Karabatsos, 2003; Reise & Due 1991). Reise and Due (1991) found that researchers need at least 20 items of varying difficultly and low discrimination to detect aberrant responding. Karabatsos (2003) supported the previous claim that longer tests are needed. Conceptually, this is unsurprising. Parameters within IRT models are supposed to be measurement invariant and a substantial amount of data are needed to reach this ambitious goal.

Karabatsos conducted a review of the detection rates of aberrant responding for 36 person-fit statistics (2003). The study found that careless and random responders were easiest to detect of all the different types of aberrant responses collapsing across prevalence of aberrant responding and test lengths. A visual inspection of the figure displaying this finding shows that most of the statistics seem to fall approximately between a 80% to a 90% detection rate, with some being better and some being less effective than this range. This result is more optimistic than previous studies. However, when considering all of the simulated data within this study (collapsing across all test lengths, types of aberrant responding, and prevalence), the detection rates decrease to approximately a range of 70 to 80% with some being larger and some being considerably lower (e.g., lower than 60%).

Many researchers report that the number of aberrant responding cases affects the detection rates of the IRT person-fit statistics (Karabatsos, 2003; Meijer & Sijtsma, 2001). Karabatsos found that a smaller prevalence of aberrant responders than 50% seem more suitable for IRT person-fit statistics (2003). This makes sense because the calculations of the person-fit statistics rely on the estimated parameters from the data. If a large portion of the sample involves aberrant responding, the estimation programs will inaccurately estimate parameters. This leads to aberrant responders fitting the IRT model (Emons, 2009).

A test taker who engages in more IER frequently is a more severe case of aberrant responding that someone who engages in IER in only a few instances. A person with more aberrant responding is easier to detect by using IRT person-fit statistics than someone with less (Emons, 2008; Meijer & Sijtsma, 2001). The reason is because the person who engages in infrequent IER would yield only a slight, possibly negligible, misfit whereas the more severe case is a clear misfit.

A review of the person-fit literature indicates that there are over forty different person-fit statistics (Meijer & Sijtsma, 2001). However, I will focus this discussion on the traditional Standardized Log-Likelihood statistic (l_z) and the Person Fluctuation Parameter (PFP) conceptualized as a person-fit statistic for the purposes of this article.

The l_z statistic is conceptually a standardized likelihood of a response pattern given the ability and item parameter estimates (Drasgow, Levine, & Williams, 1985). Therefore, larger numbers indicate better fit. The standardization of the statistic makes the result less dependent on ability estimates (Meijer & Sijtsma, 2001). This is more useful for the purposes of this article than the unstandardized version (l) as a person's underlying ability and IER behaviors should be distinct, unrelated constructs. Note that IER does not include guessing the answers. Logically, it makes sense that someone of low ability would guess at the answers which can look like inconsistent responding, a form of IER, but the process through which these response patterns appear are contradictory. A test taker cannot be guessing and engaging at IER at the same time.

In the review by Meijer and Sijtsma (2001), the authors found that l_z is an effective IER index, detecting 67% of person misfit due to aberrant responding. However, this study assumed that the true thetas were equivalent to the estimated thetas, but, regardless of the estimation procedure, they are not equivalent. I rectified this issue by using only estimated parameters in the calculation of the l_z statistic.

The other "person-fit statistic" is the PFP. Traditionally, the PFP is an alternative conceptualization of the item discrimination parameter in classical IRT. Whereas item discrimination is an item characteristic that describes variability in response for a given ability estimate centered on the difficulty parameter, the PFP is a parameter that characterizes this variability as due to the person. Lumsden (1977) proposed a model that incorporated the PFP and constrained the item discrimination parameter to be the same for each item. Because the PFP is on the same metric as the item discrimination parameter, higher parameter estimates indicate more consistent responding, and subsequently, lower estimates are an indication of inconsistent responding and/or IER behaviors within the context of the consistency indices previously discussed.

The largest disadvantage of IRT is that smaller tests and smaller sample sizes cannot be effectively examined with this method. Although this seems to be an issue with some measures, some, like personality surveys, tend to be large (e.g. MMPI and

IPIP) and they commonly administered the tests of these constructs online. This should encourage large sample sizes.

Purpose of Present Study

Past research has examined at the efficacy of the various a posteriori methods and some have looked at the efficacy of IRT person-fit statistics. However, research has yet to compare the IRT and the other methods concurrently. I believe this is due to the complexities of IRT. IRT requires a deep understanding of the methods to properly interpret the results, but the flexibility of IRT can make it a powerful tool for detecting IER on self-report measures.

I compared the IRT person-fit statistics to psychological synonyms, even-odd split and the long string methods within simulated samples with the goal of discovering the best method for detecting IER. With this goal in mind, I asked four research questions:

1. Which IER index is the best at detecting consecutive responding (i.e., 1,1,1,1, etc.)?
2. Which IER index is the best at detecting non-consecutive responses that follow a pattern (i.e., 1,5,1,5,1,5,1,5, etc.)?
3. Which IER index is the best at detecting random responding that follows a normal distribution?
4. Which IER index is the best at detecting random responding that follows a uniform distribution?

A common complaint about simulated data is that it does not reflect real world circumstances. This is arguably true, but it is impossible to assess the true accuracy of the indices without having known parameters. Therefore, I expanded previous research by including a test of the indices on a simulated sample with all known forms of IER. From this sample, I asked the following research question:

5. Which index performs the best in the mixed-IER samples?

From the results of the previous questions I asked the last research question:

6. Overall, what is the best index for detecting IER?

Method

Design

I used a fully-crossed 2x4x4 design with 100 replications for each condition plus one mixed-IER condition. The three manipulated conditions are: total number of cases, type of IER within the sample, and the prevalence of IER. The total number of cases was either be 500 or 1,000. The type of IER was manipulated using four conditions: random responding sampled from a uniform distribution, random responding from a normal distribution, a response set with the same response in succession (e.g., 1,1,1,1, etc.), and a response set with different consecutive responses but follows a pattern (e.g., 1,5,1,5,1,5, etc.). The first two conditions come from Meade and Craig (2012). The last two conditions are needed for comparing IRT to the other indices (especially the long string method). While the same response in succession has been used previously to assess the detective capabilities of indices (e.g., Meade & Craig, 2012), the repeating pattern with alternating responses condition is novel. This manipulation is needed to characterize the people that follow a pattern, but do not give the same answer for many consecutive items. Finally, I varied the percent of items within a sample with IER by 10% (10 items), 25%

(25 items), 50% (50 items), and 100% (100 items). This will cover the entire range of the possible prevalence of IER within a single case. For the mixed IER condition, I included 100 samples that have a total of 120 cases randomly sampled from the IER conditions and 880 non-IER cases.

I included the mixed IER samples because they are a better of reflection of an empirical sample. There is only one type of IER represented within the non-mixed IER samples. This probably does not reflect the true nature of an empirical sample where people engage in different types of IER (e.g., some respond randomly and some respond with response sets). Even though this analogue to a real-world sample is not a perfect representation, it is impossible to know the true number of IER cases within a real-world sample. Therefore, this simulation of a real-world sample is the best method by which researchers can assess detection rates and false positives while mimicking the properties of a real-world sample.

Data Generation

I generated the samples in the data analysis program R, from the Graded Response Model previously described. The samples had five dimensions with 20 items each rated on a 5-point graphic rating response format (100 items in all). Each sample was generated from the same priors (difficulty, discrimination, and ability parameters). The lowest difficulty parameter for each item were generated from a normal distribution with mean = -2 and SD = .45. The subsequent difficulty parameters were calculated by adding 1.3 from the previous difficulty parameter. I used a common discrimination parameter of $\alpha = 2$. All ability parameters were generated from a standard normal distribution for each person and for all five dimensions. I forced correlations among the ability parameters so that they reflected the correlations found in Table 1. These correlations make the samples mimic the intercorrelations between the subscales on the NEO-FFI found by McCrae and Costa (2004).

Each sample had a constant 12% prevalence of IER cases (i.e., 120 IER cases in the 1,000 case samples and 60 IER cases in the 500 case samples). This differs from the prevalence of IER within a single respondent. The prevalence of IER cases refers to the number of cases in which I inserted IER. The prevalence of IER within a single respondent refers to the number of item in which I inserted IER within a single case. This constant prevalence of IER comes from the estimates of the prevalence of IER cases suggested by Meade and Craig (2012). The IER was inserted into the samples via R using a random number generator for the random responding conditions. For the consecutive responses, I used a random number generator in R to select a number and inserted that number a pre-determined number of times consecutively in each dimension for the consecutive responding conditions. For the cases that use a response set that does not include using the same response for consecutive items, I used a pattern of interchanging responses between 1 and 5 for half of the IER cases and 5 and 1 in the other half.

For the mixed IER samples, I generated 880 non-IER cases. For the IER cases, I generated a separate sample that consists of IER cases only. Within the IER sample I generated 50 cases for each IER type and each within-case IER prevalence (16 conditions in total). I randomly sampled without replacement 120 cases from the IER sample to insert into the mixed IER sample. I repeated this process to make a total of 100 samples.

The longstring conditions only represented 3.33% (four cases) of the IER conditions based on Meade & Craig (2012) reporting that longstring is rarely used.

Data Analysis

Parameter estimation. After inserting IER cases into the samples, I estimated the location parameters using "mirt" package in R. These estimated item parameters are required for the IRT person-fit statistics and for theta and PFP scoring using expected a posteriori (EAP) estimation.

Using Ferrando's (2009) recommendation for estimating PFPs, the item discrimination parameters were constrained to be equal across all items. Then, I estimated the PFPs and thetas, using the following formulas:

$$\gamma_i = E(\gamma|x_i) = \frac{\int_\theta \int_\gamma \gamma L(x_i|\theta,\gamma) f(\theta) f(\gamma) d\gamma d\theta}{\int_\theta \int_\gamma L(x_i|\theta,\gamma) f(\theta) f(\gamma) d\gamma d\theta}$$

and

$$\theta_i = E(\theta|x_i) = \frac{\int_\theta \int_\gamma \theta L(x_i|\theta,\gamma) f(\theta) f(\gamma) d\gamma d\theta}{\int_\theta \int_\gamma L(x_i|\theta,\gamma) f(\theta) f(\gamma) d\gamma d\theta}$$

Where:

γ = PFP quadrature point

γ_i = PFP for each participant *i*

θ = Theta quadrature point

θ_i = Theta participant *i*

$L(x_i \mid \theta,\gamma)$ = the likelihood of a response given at each point of the theta and PFP quadrature

$f(\gamma)$ = density distribution of the PFPs

$f(\theta)$ = density distribution of the thetas

The quadrature for the PFP estimation ranged from .1 to 5 in increments of .1. I used a lognormal density function where:

$$Mean = log\left(\frac{\alpha}{\sqrt{1+\frac{\sigma_\alpha}{\alpha^2}}}\right)$$

$$SD = \sqrt{\log\left(1+\frac{\sigma_\alpha}{\alpha^2}\right)}$$

Where:

α = common item discrimination value

σ_α = standard deviation of the constrained item discrimination parameter (I set it to be one in this study)

The above equation forms the parameters needed in order to treat the PFP as normally distributed and then make all the values positive to keep it on the same metric as the item discrimination parameter. The quadrature for theta estimation will ranged from -4 to 4 in increments of .1 with a standard, normal density distribution.

IRT person fit (l_z). I chose to use two person fit statistics (l_z and the PFP described above) to assess IER detection rates. l_z is the standardized form of l_0 which involves calculation of a log- likelihood of a response pattern. Conceptually, this is a sum of the likelihood of each response given the response model and then that sum is standardized. A smaller likelihood indicates a misfit between the test taker and the model. I chose l_z as this seems to be a commonly used statistic among

Industrial/Organizational psychologists, and it has shown to be effective at detecting IER. This statistic is calculated by the following formula:

$$l_z = \frac{[l - E(l)]}{\sqrt{v(l)}}$$

Where:

$$l = \sum_{j=1}^{J} \left[X_{nj} (\ell n P_{nj1}) + (1 - X_{nj})(\ell n P_{njo}) \right]$$

$$E(l) = \sum_{j=1}^{J} \left(P_{nj1} \ell n \left[P_{nj1} \right] \right) + \left(P_{nj0} \ell n \left[P_{nj0} \right] \right)$$

$$v(l) = \sum_{j=1}^{J} \left(v_{nj} \right) \left(\ell n \frac{P_{nj1}}{P_{nj0}} \right)^2$$

Where:

J = number of test items

X_{nj} = examinee n score on the test item J

P_{nj1}= probability of that endorsing the chosen response

P_{nj0}= probability of not endorsing the chosen response, where $P_{nj0} = 1 - P_{nj1}$

v_{nj} = variance, where $v_{nj} = P_{nj1} P_{nj0}$.

Traditional IER detection indices. Below I review how to conduct some of the traditional *post hoc* statistical procedures for detecting IER.

Even-odd split. This index is essentially a split-half correlation for each person. The two halves are found by using even-numbered items for one half, and then using the odd-numbered items for the other within each unidimensional scale. Then a Pearson correlation is used to assess person reliability. If the resulting coefficient is strong this indicates that the person responded to similar items similarly. As such, a stronger positive correlation indicates more consistent responding, or less IER. A minimum requirement for this analysis is that there are 30 pairs of items. This is needed to help stabilize the correlation. When there are fewer pairs, the unstable correlation coefficient renders this index uninterpretable (e.g. it is difficult to interpret if there is a difference between $r = 0.30$ and $r = 0.50$). As such, when a correlation coefficient is unstable, introducing more data can drastically change the coefficient.

Psychological synonyms. The only difference between this index and the even-odd split index lies in how the item pairs are chosen. In psychological synonyms, the pairs are chosen by inter-item correlations with a coefficient of at least $r = 0.60$. However, in this study, I had to relax this to a correlation of $r = 0.45$ in order to consistently achieve the number of pairs needed to run the analysis effectively. The pairs are divided into two halves to run a split-half correlation. Smaller and all negative correlations indicate IER. The minimum requirement of 30 pairs of items to calculate the split-half correlation is needed here.

Long string. This statistic is the largest number of times a single response appears consecutively. For example, someone could have answered with a "1" ten consecutive times. In that same response pattern, that person could have answered with a "3" five consecutive times. The long string statistic would be ten in the above example because "1" appeared consecutively more times than the "3".

Empirical cut-offs. None of the proposed indices have a well-established cut-off score for use in categorizing a case into an IER or non-IER case. One way to establish cut-offs involves converting the index into a percentile rank and then testing various theoretically plausible ranks to determine which has the highest detection rates (power) with the lowest error rate. However, the issue with this approach is that if the theoretical cut-off is incorrect, this could affect the power rates. Therefore, instead of having a single cut-off for each index, I used six percentile rank bands that cover the entire range of possibilities. The bands included: 0% - 6%, 6.1% - 12%, 12.1% - 50%, 50.1% - 87.9%, 88% - 93.9%, and 94% - 100%. The bottom and top two bands come from the 12% prevalence of IER cases. The middle two are used just to cover all possibilities, and were not used in the results.

Power. Power is a measure of the detecting rates of the indices by giving the proportion of true positives detected. The calculation is performed with the following formula:

$$Power = \frac{\textit{Number of true positives detected}}{\textit{Total number of true positives}}$$

A larger proportion indicates better effectiveness. I calculated a power rate for each index, for each percentile rank band, and for each condition.

Error. Error is another measure of the accuracy of each IER indicator. Error measures the proportion of false positives. Error is calculated by the following formula:

$$Error = \frac{\textit{Number of false positives detected}}{\textit{Total number of true negatives}}$$

A small proportion indicates low error. I calculated an error rate for each index, for each percentile rank band, and for each condition.

Results

Within all indices, the power rates increased and the error rates decreased as the amount of IER within the IER cases increased excluding the non-consecutive longstring index. This is unsurprising as index accuracy should improve for more egregious cases of IER.

RQ 1. *Which IER index is the best at detecting consecutive responding (i.e., 1,1,1,1, etc.)* The results for this question can be found on Table 2. Regardless of sample size, the best index to use to detect consecutive responding for the 10% and 25% conditions is the l_z person-fit statistic. Regardless of sample size, in the 10% IER conditions the average power rate was .44 with an average error rate of .08. In the 25% IER conditions the average power rates only ranged from .69 in the *n* = 500 conditions and .70 in the *n* = 1,000 conditions. The average error rates stayed a constant .04. In the 10% IER condition, the PFP performed just as well as l_z in the *n* = 500 conditions (average power estimate of .44 and average error rate of .08). However, the PFP does not perform as well as the l_z in the 25% IER condition with an average power rate ranging from .56 for the *n* = 500 conditions to .58 for the *n* = 1,000 conditions. The average error rates were also larger (.06 for both sample size conditions). However, as convention dictates, a power estimate of 0.8 is needed in order for the power to be considered acceptable. Therefore, none of the indices were acceptable detectors of IER in the 10% and 25% IER conditions. Regardless of sample size, in the 50% and 100% IER conditions, the longstring index reached this convention of acceptable power (0.82 and 1.0, respectively with average error rates of .04 and 0). Therefore, it outperformed all of the other indices. This makes sense as the longstring index is calculated by the number of consecutive responses (i.e., it is measuring itself).

All indices were able to detect IER cases perfectly (1.0 power rate and 0 errors) in the 100% IER conditions. However, all indices (excluding the longstring index) required using cut-offs that are theoretically incorrect to use. For example, in the non-consecutive index, the smallest numbers were indicators of IER although large numbers should indicate IER. This complicates the interpretation of these indices, rendering them useless by themselves.

RQ 2. *Which IER index is the best at detecting non-consecutive responses that follow a pattern (i.e., 1,5,1,5,1,5,1,5, etc.)?* The results for this question can be seen in Table 3. Regardless of sample size, the best overall index for detecting non-consecutive, patterned responses was the PFP. It produced no error and perfect power rates (1.0) for the 25% to 100% IER conditions regardless of sample size. The weakest power rates the PFP produced came from the 10% IER conditions, and they still reached the convention of acceptable power (.88 for the n = 1,000 conditions and .89 for the n = 500 conditions with a constant average error rate of .01).

The l_z statistic performed well in the 25% and 50% IER conditions with no error with perfect and almost perfect (.99) average power rates. In the 10% and 100% IER conditions, the l_z did not perform as well. It did not reach acceptable power in the 10% IER conditions with a constant average power rate of .67. In the 100% IER conditions, larger numbers indicated IER, which is anti-theoretical. This renders the l_z useless in detecting non-consecutive patterned responses when all response options are answered in a pattern.

The psychological synonyms index performed well in the 50% IER condition, with an average power rate of .93 in the n = 500 conditions and .95 in the n = 1,000 conditions with a constant average error rate of .01. However, the psychological synonym index did not perform as well as the l_z and PFP statistic in the 50% IER conditions. It also performed perfectly in the 100% IER condition (power of 1.0 with no error).

The non-consecutive longstring index produced perfect power rates in the 50% and 100% IER conditions. It produced and almost no error (.01) in the 50% IER conditions and no error in the 100% IER conditions. However, because it produced some error in the 50% IER conditions, the PFP performed better. It performed dreadfully in the 10% and 25% IER conditions. The largest average power rate in this condition was .04. Overall, I would suggest using the PFP to detect non-consecutive, patterned responding.

RQ 3. *Which IER index is the best at detecting random responding that follows a normal distribution?* Results for this table can be found on Table 4. All indices, excluding the longstring indices, tended to improve detecting random responding from a normal distribution as sample size increased although this was not always true. When there were decrements, the average power rates decreased by only .01, which I would argue is not practically relevant. There were only three practically relevant improvements. (1) In the 50% IER conditions, the average power rate increased from .26 to .35 for the psychological synonyms. (2) In the 100% IER conditions, the average power rate increased from .81 to .85 in the psychological synonyms. (3) In the 10% IER conditions, the average power rate increased from .30 to .35 for l_z. Error rates almost always remained equivalent.

The only acceptable power rates came from the psychological synonyms, the PFP and the l_z. The psychological synonyms only reached acceptable average power rates in the 100% IER conditions (.81 in the n = 500 conditions to .85 in the n = 1,000 conditions). The average error rates were .03 and .02 respectively.

The PFP had acceptable average power rates in the 50% and 100% IER conditions. In the 50% IER conditions there was a minute decrement from an average power rate of .88 to .87 as sample size increased. In the 100% IER conditions, the average power rate was a constant .96. Although they did not reach acceptable power rates, in the 10% IER conditions, the PFP performed the best with the highest average power rates (.36 for n = 500 conditions and .38 for n = 1,000). That said, the error rates, although constant, were fairly large (.09).

The l_z slightly outperformed the PFP in the 50% and 100% IER conditions with average power rates of .90 and .98 respectively and had no error in the 100% IER conditions. The average power rates were unaffected by sample size. Also, l_z did not perform as well as the PFP in the 10% IER conditions, but it was as closest to the PFP with average power rates of .30 for the n = 500 conditions and .35 for the n = 1,000 conditions. Overall, I believe that l_z is the best detector of random responding following a normal distribution, but followed closely by the PFP.

RQ 4. *Which IER index is the best at detecting random responding that follows a uniform distribution?* Results for this question can be found in Table 5. Unlike in Table 4, there were not any drastic changes in average power rates due to sample size, but there was one notable change in average error rates. In the 25% IER condition for the PFP, the error rate increased from .02 to .05 as sample size increased.

The only acceptable power rates came from the PFP and l_z. The PFP had acceptable average power rates in the 25% through 100% IER conditions for both sample size conditions. The power rates slightly changed as sample size increased although these changes were minute. In the 25% IER condition, the average power rates decreased from .84 to .83. In the 50% IER condition, the average power rates decreased from .99 to .98. Regardless of sample size, the PFP had perfect average power rates with no error in the 100% IER condition. The PFP performed the best in the 10% IER condition although the power rates did not reach .80 and the average error rate was high (.08 for n = 500, and .07 for n = 1,000). The average power rate increased from .48 to .49 as sample size increased.

Similar to the PFP, the l_z had acceptable average power rates in the 25% through 100% IER conditions for both sample size conditions. The power rates had one minute change as sample size increased. In the 25% IER condition, the average power rates decreased from .83 to .82. In the 50% IER condition, the average power rates were a constant .99. In the 100% IER condition, l_z had perfect average power rates with no error. The l_z did not perform as well as the PFP in the 10% IER condition, but it was close. The average power rate was a constant .44 as sample size increased with an average error rate of .08. Overall, I would consider both the l_z and PFP to be equal in detecting random responding from a uniform distribution

RQ 5. *Which index performs the best in the mixed-IER samples?* The results to answer this question can be found in Table 6. In the mixed-IER samples, the l_z and the PFP performed similarly. The l_z performed the best with an average power rate of .76 and an average error rate of .03. The PFP had an average power rate of .75 with an average error rate of .03. The smallest power rate of both of IRT indices were larger than that of every other indices maximum power rate. Overall, the IRT indices had lower error rates than every other index as well. The only discrepancy is that the average and minimum error rates of the non-consecutive longstring index were comparable to the maximum and average error rates of the IRT indices.

It is also noteworthy that none of the indices reached acceptable power rates with the exception of some instances of the IRT indices reaching 0.80 and larger. This is probably due to the prevalence of the 10% IER conditions within the samples. In almost every condition, the 10% IER cases were difficult to detect and therefore, probably attenuated the power rates.

RQ 6. *Overall, what is the best index for detecting IER?* In this study, there was not a single, clear winner for the best index for detecting IER. However, the l_z and the PFP performed the best overall. The PFP index performed well in comparison to all other indices, excluding l_z, in all conditions except the 100% longstring conditions. The l_z index performed well in comparison to all other indices, excluding the PFP, in all but the 100% longstring and the 100% non-consecutive longstring conditions. The PFP out-performed l_z in a few conditions, but l_z out-performed the PFP in a few conditions as well. Therefore, I believe that both the PFP and the l_z are equally useful for detecting IER.

Discussion

This study was simulation study which extended the previous work done by Meade and Craig (2012) and Huang et al. (2012). I examined the power and error rates of many of the traditional *post hoc* indices in conjunction with two IRT approaches (l_z

and PFP) to detecting IER. Overall, the IRT indices outperformed all other indices and performed almost equally to each other. There were very few differences between the performances of these indices due to sample size. In the mixed-IER samples, the IRT indices performed the best. Given these results, I conclude that the IRT approach provides the best indices for detecting IER accurately.

Although the IRT indices clearly out-performed the other indices, they are not without limitations. IRT person-fit statistics require many items for one scale in order to perform well (Reise & Due, 1991). That said, Reise and Due (1991) did not include the PFP in their study. It is possible that the constraint of requiring 20 items might not hold as strongly on the PFP as it does for l_z because the PFP is an IRT parameter instead of a person-fit statistic. This claim has yet to be tested, but it could educate researchers further on the abilities of the PFP in detecting IER. That said, scales with only a few items (e.g., 3 to 8 items) would most likely not yield stable results as the PFP estimation procedure (and IRT approaches in general) tend to require more items. Future research should attempt to find the minimum number of items needed in order to yield an accurate PFP estimate.

Another limitation of the use of IRT indices is that they require a large sample size. I chose $n = 500$ and $n = 1,000$ for our sample sizes and even 500 cases can be difficult to achieve. However, it is interesting that I found only a few meaningful differences within the IRT indices power rates between the two sample sizes. Also, the few meaningful differences I did find, were in the 10% of items with IER conditions where the power rates were low anyway. This leads me to suspect that $n = 500$ might not be the smallest sample size in which IRT can be used to detect IER. Future researchers should address this concern with sample sizes in between 250 and 500 people.

For the non-consecutive patterned responding conditions, I did not address all possible combinations of patterns. However, this would a huge undertaking as there are numerous possibilities for patterned responses. The number of these patterns exponentially grow as the number of items and response options increase. Given the low prevalence of this behavior and the amount of work it would take, I believe that this would not provide much to the IER literature. However, manipulating the response options to be more similar (like a repeating pattern of 1 and 2, or 4 and 5) might be feasible. Although the IRT indices should be able to detect this, it depends on the variation between the location parameters across the items. If the location parameters were similar, it might seem that the person would fit the model. However, if the location parameters were different, the person could be seen as not fitting the model.

Manipulating the sample size of the mixed IER samples would have been useful. Because I sampled 120 cases (four of which came from the eight longstring conditions), it is likely that each condition was semi-represented equally. If I were to have simulated samples with $n = 500$ (60 IER cases), it would be more likely that the power rates would vary more due to under and over representation of the IER conditions. This could help researchers establish boundary conditions or help researchers to interpret the validity of these indices in detecting IER.

Finally, these findings should be cross-validated in a real-world sample. Although researchers developed IRT to explain how someone may respond to an item, the samples presented in this study are only simulations. I am more confident in externalizing these findings with the addition of the mixed-IER samples than with only the single condition samples. However, these samples are still simulations, and it is impossible to simulate every response process underlying responses to a scale. Therefore, there are issues with generalizing these results to the real world.

In validating IRT IER detection protocols in a real world sample, a vital step is to ensure that the people are not fitting the model due to IER and not due other factors such as true outliers where they might be detected as an IER case, but respond with accurate information. This can be done by examining the correlations of the IRT approaches with a combination of all the traditional approaches. Another useful approach to validating IRT measures of IER is to use an approach like that used by Meade and Craig (2012), which involved a Latent Profile Analysis and then a discriminant function analysis to see if the IRT indices are able to predict classification from the Latent Profile Analysis.

Conclusions and Recommendations

I conclude that the PFP and l_z indices are great indicators of IER and are better than the traditional *post hoc* indices in large tests and sample sizes. Even though the IRT indices performed well in every condition, the long string method was the best for detecting consecutive responding by far. The issue with the long string method is that the less egregious cases are hard to detect. Therefore, I recommend that future users of IER indices use the longstring method to detect obvious cases (i.e., 50% and 100% of the items responded to with the same response) and then use the PFP and l_z together to detect the rest.

References

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI—A. *Journal of Personality Assessment, 68*, 139-151.

Beach, D. A. (1989) Identifying the random responder. *The Journal of Psychology: Interdisciplinary and Applied, 123,*101-103.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224-247.

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous items scores. *Applied Psychological Measurement*, *33*, 599-619.

Ferrando, P. J. (2009). A graded response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology*, *62*, 641-662.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75*, 581-595.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99-114

Johnson, J.A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*, 103-129.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, *1*, 477–482.

McCrae, R. R. & Costa, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, *36*, 587-596.

Meade, A. W., & Craig, S. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*, 3-8.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*, 72-87.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*(3), 217-226.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents?. *Applied Psychological Measurement*, *9*, 367-373.

Spector, P. E., & Brannick, M. T. (2009). Common method variance or measurement bias? The problem and possible solutions. In D. A. Buchanan, A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 346-362). Thousand Oaks, CA: Sage Publications Ltd.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, *95*, 334-344.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349-364.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186-191.

Figure 1.

Plot of the probabilities of endorsement for three items in Item Response Theory.

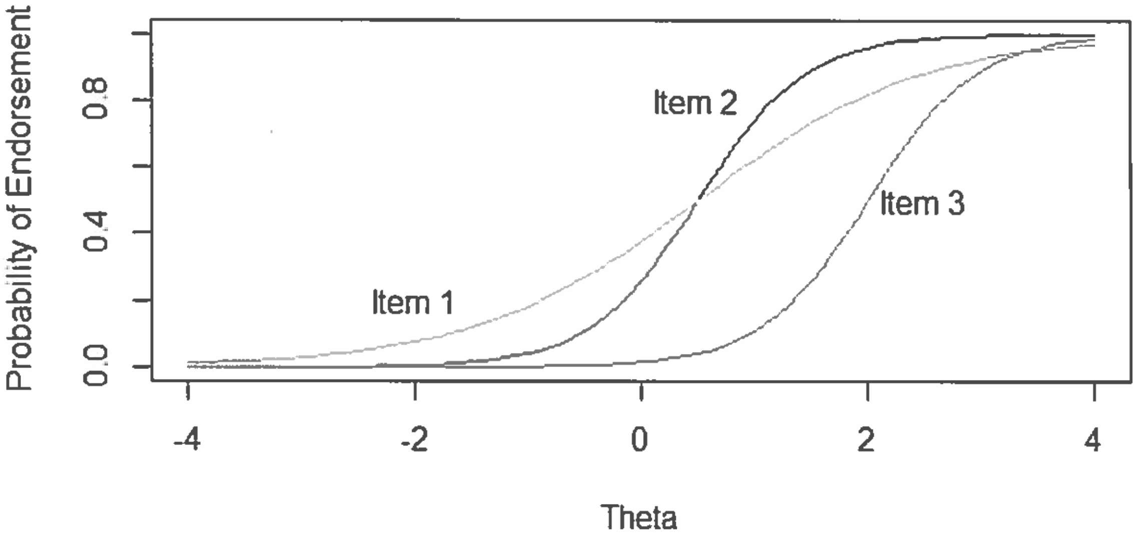

Table 1.

Intercorrelations between dimensions in each sample.

Factor	1	2	3	4	5
1	-				
2	-.38	-			
3	-.04	.24	-		
4	-.29	.22	.10	-	
5	-.42	.32	.01	.17	-

Table 2.

Results for all indices in the longstring conditions.

IER Condition		n = 500					
		Long	NC Long	Syn	Split	Lz	PFP
Longstring	10%	.11(.11)	.05(.06)	.13(.12)	.13(.12)	.44(.08)	.44(.08)
	25%	.23(.09)	.04(.06)*	.06(.13)*	.15(.12)	.69(.04)	.56(.06)
	50%	.82(.04)	.02(.06)*	.00(.14)*	.04(.13)*	.63(.05)	.48(.07)
	100%	1.0(.00)	.00(.06)*	.00(.14)*	.00(.14)*	.00(.14)*	.00(.14)*
		n = 1,000					
		Long	NC Long	Syn	Split	Lz	PFP
Longstring	10%	.11(.11)	.05(.06)	.13(.12)	.14(.12)	.44(.08)	.33(.08)
	25%	.23(.09)	.04(.06)*	.04(.13)*	.15(.12)	.70(.04)	.58(.06)
	50%	.84(.04)	.02(.06)*	.00(.14)*	.04(.13)*	.64(.05)	.48(.07)
	100%	1.0(.00)	.00(.06)*	.00(.14)*	.00(.14)*	.00(.14)*	.00(.14)*

Note. Long = Longstring Index; NC Long= Non-Consecutive Longstring index; Syn= Psychological Synonyms Index; Split= Split-Half Method Index; Lz = l_z person-fit index; PFP = Person Fluctuation Parameter; * = the best power estimates were in anti-theoretical percentile ranks; The numbers in parentheses are the error rates.

Table 3.

Results for all indices in the non-consecutive longstring condition.

IER Condition		n = 500					
		Long	NC Long	Syn	Split	Lz	PFP
	10%	.10(.12)*	.04(.06)	.11(.12)*	.14(.12)	.67(.04)	.89(.01)
Non-Consecutive Longstring	25%	.08(.12)*	.03(.06)	.00(.11)*	.17(.12)	.99(.00)	1.0(.00)
	50%	.04(.13)*	1.0(.01)	.00(.01)*	.24(.10)	1.0(.00)	1.0(.00)
	100%	.00(.15)*	1.0(.00)	.00(.00)*	.00(.14)*	.00(.14)*	1.0(.00)
		n = 1,000					
		Long	NC Long	Syn	Split	Lz	PFP
	10%	.09(.11)*	.04(.06)	.11(.12)*	.12(.12)	.67(.04)	.88(.01)
Non-Consecutive Longstring	25%	.08(.12)*	.03(.06)	.00(.10)*	.17(.12)	.99(.00)	1.0(.00)
	50%	.04(.13)*	1.0(.01)	.00(.01)*	.24(.10)	1.0(.00)	1.0(.00)
	100%	.00(.14)*	1.0(.00)	.00(.00)*	.00(.14)*	.00(.14)*	1.0(.00)

Note. Long = Longstring Index; NC Long= Non-Consecutive Longstring index; Syn= Psychological Synonyms Index; Split= Split-Half Method Index; Lz = l_z person-fit index; PFP = Person Fluctuation Parameter; * = the best power estimates were in anti-theoretical percentile ranks; The numbers in parentheses are the error rates.

Table 4.

Results for all indices in the random responding following a normal distribution conditions

IER Condition		n = 500					
		Long	NC Long	Syn	Split	Lz	PFP
Random Responding from Normal Distribution	10%	.11(.11)*	.06(.06)	.13(.12)	.16(.12)	.30(.09)	.36(.09)
	25%	.08(.12)*	.05(.06)	.20(.10)	.24(.10)	.66(.05)	.66(.05)
	50%	.06(.12)*	.04(.06)	.26(.08)	.43(.07)	.90(.01)	.88(.02)
	100%	.02(.14)*	.02(.06)*	.81(.03)	.74(.04)	.98(.00)	.96(.01)
		n = 1,000					
		Long	NC Long	Syn	Split	Lz	PFP
Random Responding from Normal Distribution	10%	.10(.12)*	.06(.06)	.14(.12)	.16(.12)	.35(.09)	.38(.09)
	25%	.09(.12)*	.05(.06)	.19(.12)	.24(.10)	.66(.05)	.65(.05)
	50%	.06(.12)*	.04(.06)	.35(.09)	.43(.07)	.90(.01)	.87(.02)
	100%	.03(.13)*	.02(.06)*	.85(.02)	.75(.04)	.98(.00)	.96(.01)

Note. Long = Longstring Index; NC Long= Non-Consecutive Longstring index; Syn= Psychological Synonyms Index; Split= Split-Half Method Index; Lz = l_z person-fit index; PFP = Person Fluctuation Parameter; * = the best power estimates were in anti-theoretical percentile ranks; The numbers in parentheses are the error rates.

Table 5.

Results for all indices in the random responding following a uniform distribution conditions

IER Condition		n = 500					
		Long	NC Long	Syn	Split	Lz	PFP
Random Responding from Uniform Distribution	10%	.10(.11)*	.05(.06)	.14(.12)	.17(.12)	.44(.08)	.48(.08)
	25%	.08(.12)*	.04(.06)	.17(.12)	.28(.10)	.83(.02)	.84(.02)
	50%	.04(.12)*	.03(.06)	.29(.10)	.48(.07)	.99(.00)	.99(.00)
	100%	.00(.14)*	.01(.06)*	.65(.05)	.75(.04)	1.0(.00)	1.0(.00)
		n = 1,000					
		Long	NC Long	Syn	Split	Lz	PFP
Random Responding from Uniform Distribution	10%	.10(.12)*	.05(.06)	.13(.12)	.17(.12)	.44(.08)	.49(.07)
	25%	.08(.12)*	.04(.06)	.16(.12)	.27(.10)	.82(.02)	.83(.05)
	50%	.05(.13)*	.03(.06)	.26(.10)	.49(.07)	.99(.00)	.98(.00)
	100%	.00(.14)*	.01(.06)*	.67(.05)	.75(.04)	1.0(.00)	1.0(.00)

Note. Long = Longstring Index; NC Long= Non-Consecutive Longstring index; Syn = Psychological Synonyms Index; Split= Split-Half Method Index; Lz = l_z person-fit index; PFP = Person Fluctuation Parameter; * = the best power estimates were in anti-theoretical percentile ranks; The numbers in parentheses are the error rates.

Table 6.

Power and error estimates of each index in the mixed-IER condition.

	Long	NC Long	Syn	Split	Lz	PFP
Average	.08(.13)	.04(.05)	.43(.08)	.41(.08)	.76(.03)	.75(.03)
Minimum	.00(.07)	.01(.04)	.29(.06)	.24(.06)	.66(.02)	.65(.03)
Maximum	.23(.22)	.09(.08)	.58(.10)	.55(.10)	.88(.05)	.82(.05)

Note. Long = Longstring Index; NC Long= Non-Consecutive Longstring index; Syn= Psychological Synonyms Index; Split= Split-Half Method Index; Lz = l_z person-fit index; PFP = Person Fluctuation Parameter. Numbers in parentheses are error rates.