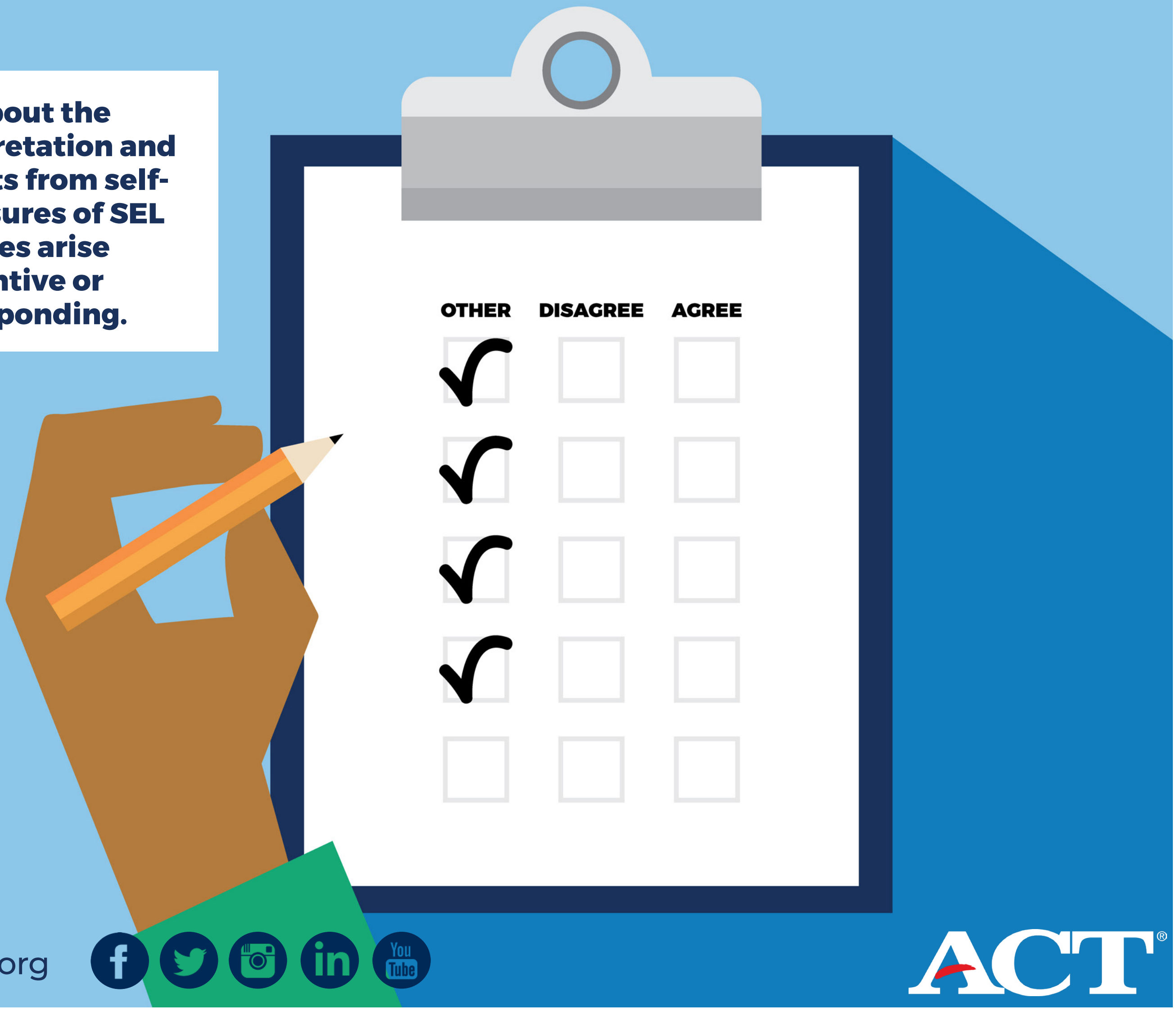
Research Report
2018-5

Detecting Inattentive Responding on a Psychosocial Measure of College Readiness

JEFFREY STEEDLE, PHD

Concerns about the valid interpretation and use of results from self-report measures of SEL competencies arise from inattentive or careless responding.



AUTHOR

JEFFREY STEEDLE, PHD

Jeffrey Steedle is a senior research scientist in Validity and Efficacy Research specializing in postsecondary outcomes research and validity evidence for ACT's workforce assessment programs.

SUMMARY

Contemporary definitions of college and career readiness extend well beyond the mathematics and English language arts skills needed for success in postsecondary courses or job training. Among other aspects of readiness, those definitions include psychosocial or social and emotional learning (SEL) competencies reflecting behaviors and attitudes that support learning and persistence.

In this study, insufficient effort responding detection methods were applied in a new and important context: self-report assessments of SEL competencies related to college readiness. This study also introduces three Engage methods and evaluated their use when reporting Engage results.

SO WHAT?

Concerns about the valid interpretation and use of scores from self-report measures stem from problems that can occur during the response process. For example, respondents may interpret items or response options differently, they may lack the insight or information needed to respond accurately, or they may exhibit biases such as socially desirable responding (Duckworth & Yeager, 2015). These threats to validity require that respondents are attentive to item content; a more fundamental threat arises from responding in a careless, random, or inattentive manner that disregards item content (Huang, Curran, Keeney, Poposki, & DeShon, 2012). This phenomenon, referred to as insufficient effort responding (IER), is thought to be associated with respondent interest, survey length, researcher-respondent interactions, and environmental distraction.

NOW WHAT?

This study generated estimates of IER prevalence similar to those from other contexts, and it illustrated the difficulties inherent in estimating IER prevalence. Results corroborated prior studies in which the effects of IER on validity evidence were small. Even so, researchers and test developers should be attentive to IER and flag individual results suspected of IER. Test users should be skeptical when interpreting certain self-report assessment results, especially those flagged by multiple IER detection methods and inconsistent with other indicators of SEL competencies.

Abstract

Self-report inventories are commonly administered to measure social and emotional learning competencies related to college readiness. If students respond inattentively or dishonestly, validity will suffer. This study applies several methods of detecting insufficient effort responding (IER) to data from ACT® Engage®. Different methods indicated that between 0.8% and 20.3% of respondents exhibited IER, but filtering those students from the data resulted in negligible improvements in criterion-related validity, coefficient alpha, convergent validity, and confirmatory factor analysis model-data fit. Even so, researchers are advised to investigate IER. Analyses affirmed that the IER detection methods effectively flagged suspect item score patterns, so these methods may still be used to flag individual results as potentially invalid.

Table of Contents

Detecting Inattentive Responding on a Psychosocial Measure of College Readiness

Jeffrey Steedle, PhD

Introduction

Contemporary definitions of college and career readiness extend well beyond the mathematics and English language arts skills needed for success in postsecondary courses or job training. Among other aspects of readiness, those definitions include psychosocial or social and emotional learning (SEL) competencies reflecting behaviors and attitudes that support learning and persistence. For example, Conley's (2007) operational definition of college readiness includes academic behaviors such as self-awareness, self-monitoring, and self-control in academic pursuits. Likewise, the behavioral skills component of the ACT Holistic Framework™ of Education and Work Readiness includes attributes related to self-confidence, flexibility, and perspective taking (Camara, O'Connor, Mattern, & Hansen, 2015). Even some state definitions of college and career readiness include factors like collaboration, communication skills, and resilience (Mishkind, 2014). SEL competencies are desirable products of education, and they are known to explain variation in postsecondary outcomes in addition to that explained by prior academic achievement (Robbins, Lauver, Le, Davis, & Langley, 2004; National Research Council, 2012).

Enthusiasm for measuring SEL has followed from acknowledgment of its role in student success (Naemi, Burrus, Kyllonen, & Roberts, 2012; Levin, 2013). Self-report questionnaires are a common and convenient method for gathering information about SEL competencies. Yet, as Herman and Hilton (2017) reported, most current assessments are "uneven in quality, providing only limited evidence to date that they meet professional standards of reliability, validity, and fairness" (p. 8). Given that most assessments have yet to provide strong evidence that they meet professional measurement standards (AERA, APA, & NCME, 2014), their use in high-stakes contexts cannot be supported, and even their use in low-stakes contexts may be problematic.

In part, concerns about the valid interpretation and use of scores from self-report measures stem from problems that can occur during the response process. For example, respondents may interpret items or response options differently, they may lack the insight or information needed to respond accurately, or they may exhibit biases such as socially desirable responding (Duckworth & Yeager, 2015). These threats to validity require that respondents are attentive to item content; a more fundamental threat arises from responding in a careless, random, or inattentive manner that disregards item content (Huang, Curran, Keeney, Poposki, & DeShon, 2012). This phenomenon, referred to as insufficient effort responding (IER), is thought to be associated with respondent interest, survey length, researcher-respondent interactions, and environmental distraction (Meade & Craig, 2012). IER may be exhibited by short response times or by response patterns reflecting lack of internal consistency, high variability, excessive repetition, or failure to notice negatively-worded items or items designed to check respondent attention.

Besides having consequences for the interpretation and use of individual scores, IER in the aggregate can potentially impact validity evidence for self-report measures. If IER is random, it would be expected to introduce measurement error that deflates criterion-related validity coefficients, reduces reliability, and distorts factor structure (McGrath, Mitchell, & Kim, 2010). However, real-life respondents rarely behave randomly, and the effects of non-random IER on validity evidence are unpredictable (Meade & Craig, 2012). Numerous statistical methods have been developed to detect IER (Curran, 2016), yet the use of those methods is rarely reported, even in top-tier journals (Ran, Liu, Marchiondo, & Huan, 2015). As a first step in validity studies, detection methods could estimate the prevalence of IER in the testing population. In addition, the removal of data reflecting IER could help ensure that validity studies provide trustworthy results.

The major objective of this study was to examine the effects of IER on validity evidence in an emergent field of measurement: assessing SEL competencies associated with college readiness. For this study, several methods of detecting IER were applied to data from ACT® Engage®, a self-report inventory including 108 Likert-scale items divided between 10 subscales (ACT, 2016). Results estimated the prevalence of IER for high school students and the associations between different IER indices.

The effects of IER on validity evidence were examined by removing data reflecting apparent IER and observing changes in criterion-related validity coefficients, subscale reliability, correlations among subscales, and factor analysis model-data fit. In all, results estimate IER prevalence and the extent to which validity evidence might be biased by IER, thereby providing guidance for future validity studies and score reporting for measures of SEL competencies related to college readiness.

Background

This section provides a review of methods for detecting IER and a summary of prior research examining the effects of IER on validity evidence. Only methods of IER detection applicable to Engage data are reviewed here; other methods require response-time data or attention-check items embedded in the assessment (Curran, 2016). In the following descriptions, *raw* response data refers to the item scores as they were captured, and *rescored* response data refers to the item scores after reversing the scoring of negatively-worded items. Most IER detection methods are appropriate for a certain type of response data (raw or rescored).

IER Detection

Engage methods.

ACT currently implements three methods for detecting IER in Engage data (ACT, 2016). Respondents flagged by these analyses are indicated in the "advisor report" provided to the test administrator. The first method identifies respondents who do not apparently distinguish between positively-worded and negatively-worded items. Specifically, a respondent is flagged if the mean absolute difference between scores on positively-worded and negatively-worded items in the rescored data is greater than or equal to 2.0 (for items scored from 1 to 6). This method assumes that conscientious respondents should endorse options reflecting a consistent level of the measured construct, which would result in similar means for positively- and negatively-worded items. For example, a respondent who chooses the fifth option for all items would have a positively-worded mean of 5.0, a negatively-worded mean of 2.0, and a mean absolute difference of 3.0. Like some other approaches, this method treats the assessment as a whole, rather than a series of subscales, so it relies on the fact that items from different subscales correlate positively.

The other two Engage methods focus on detecting respondents with invariant item scores. Specifically, respondents are flagged if the standard deviations of their item scores are 0.50 or less or if they have a certain item score on 90% or more of the items to which they responded. These approaches assume that conscientious respondents should exhibit a certain degree of item score variability resulting from intra-individual differences across items and subscales. Respondents are flagged if they select response options reflecting the same or similar level of the measured constructs for most items. For example, a respondent choosing the most socially desirable response to every item would be flagged due to a standard deviation of 0 and for having the same item score on 100% of items.

Item score variance.

The standard deviation approach used for Engage is most similar to intra-individual response variability (IRV), which is also the standard deviation of item scores, but calculated on raw item scores (Dunn, Heggestad, Shanock, & Theilgard, 2016). By flagging respondents with low IRV, this approach catches respondents with long strings of the same response. This method assumes that conscientious respondents should exhibit response variability because their levels differ on various constructs measured by an assessment and because scores should vary across positvely- and negatively-worded items. Consistent with expectations, respondents flagged for low IRV on a personality inventory exhibited lower conscientiousness and higher proneness to boredom (Dunn et al., 2016).

The Engage standard deviation approach is also akin to the inter-item standard deviation (ISD; Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2015), which has also been called intra-individual variance (Baumeister & Tice, 1988). The assumption behind ISD is that conscientious respondents should consistently endorse response options reflecting a similar level of the measured construct, which would result in a low ISD in the rescored data. Thus, respondents with high ISD may exhibit IER. This contrasts with the Engage method of flagging respondents with low item standard deviations, but the goals of these approaches differ. ISD is intended to detect random or inconsistent responding behavior, whereas the Engage approach detects excessively consistent responding.

In prior research, simulated random respondents exhibited high ISD (Marjanovic et al., 2015), but truly random responding is unlikely to be observed in real data. Indeed, even respondents instructed to answer quickly without thinking sometimes produced response patterns that looked like conscientious responders (Huang et al., 2012). Moreover, high ISD is supposed to indicate IER, but ISD was positively associated with a measure of conscientiousness (Austin, Deary, Gibson, McGregor, & Dent, 1998). This result suggests that conscientious respondents may sometimes exhibit greater response variability because they discriminate carefully between response options. Given these research findings, some combination of flagging students with very low or very high item standard deviations may be worthy of consideration.

Long-string analysis. The long-string approach identifies respondents with unusually long sequences of the same response in the raw data. Such sequences are assumed to reflect IER, especially when negatively-worded items are present. With this approach, a normative cutoff is established for each response option, and respondents get flagged if any of their maximum response string lengths exceed the corresponding cutoff. For example, Costa and McCrae (2008) examined five-category personality inventory items and found that conscientious responders never chose the same option more than 6, 9, 10, 14, and 9 times for strongly disagree, disagree, neutral, agree, and strongly agree, respectively. As expected, long-string analysis was effective for identifying respondents who select the same response for many consecutive items (Meade & Craig, 2012). However, another study indicated that long-string analysis was less sensitive than other methods when identifying respondents assigned to respond without effort (Huang et al., 2012). Perhaps few respondents approached their task in a manner resulting in long strings, but authentic IER could manifest differently.

Individual consistency. Individual consistency approaches attempt to identify respondents with inconsistent scores on items for which similar scores would be expected. In the individual reliability (or even-odd consistency) approach, for example, each subscale is split in two (randomly or even/odd items). After rescoring the negatively-worded items, scores on the first halves are correlated with scores on the second halves, and those correlations are adjusted using the Spearman-Brown formula to account for unreliability associated with using shorter tests (Jackson, 1976). Assuming that conscientious respondents should score similarly on both halves of each subscale, low individual reliability may indicate IER. Because individual reliability may be highly dependent on the items in a particular subscale split, Curran (2016) proposed resampled individual reliability, which is the average individual reliability over many random subscale splits.

In the psychometric antonyms approach (Goldberg & Kilkowski, 1985), the correlations among all items are calculated, and the 30 item pairs with the strongest negative correlations are identified (often pairs of positively- and negatively-worded items). Then, for each respondent, the correlation between scores on those 30 items is calculated. This value should be strongly negative for respondents with similar scores across items in the rescored data. In research, the sign of the psychometric antonym index is often reversed to make it correlate positively with other indices for which low values indicate IER.

In prior research, individual reliability and the psychometric antonyms index were highly correlated (*r* = .69, *p* < .001) and moderately sensitive when detecting respondents who were assigned to respond without effort (Huang et al., 2012). Other studies provide additional evidence of convergent validity for these indices. In one study, exploratory factor analysis indicated that individual reliability, psychometric antonyms, and Mahalanobis distance (described below) loaded strongly on the same factor (Meade & Craig, 2012). Another study showed that these indices were strongly correlated with scores on an “infrequency” scale consisting of items on which all attentive respondents should provide the same responses (e.g., “I work twenty-eight hours in a typical work day”; Huang, Bowling, Liu, & Li, 2015).

Aberrant score patterns. Another class of methods for detecting IER involves examining entire response patterns for significant deviations from expectations. One such measure, the Mahalanobis distance (Mahalanobis, 1936), may be used to detect multivariate outliers. This index quantifies the distance between an individual’s item score pattern and the mean pattern in a *J*-dimensional space, where *J* is the number of items on an instrument. Specifically, Mahalanobis distance (*D*) is calculated as $\sqrt{(x_i - \bar{x})C_x^{-1}(x_i - \bar{x})^T}$, where x_i is respondent i's vector of item scores, $\bar{x}$ is the vector of mean item scores, and C_x is the covariance matrix for all items. A respondent is flagged when the squared Mahalanobis distance is greater than a critical chi-squared value with *J* degrees of freedom (i.e., $D^2 > \chi^2_{J,\alpha}$).

In simulation research, Mahalanobis distance was sensitive to extreme and random responding, but not to socially-desirable faking (Zijlstra, van der Ark, & Sijtsma, 2011). In another study, Mahalanobis distance was the best method for detecting simulated inattentive respondents with 25% of items having random item scores drawn from a uniform distribution (Meade & Craig, 2012). However, when item scores were drawn from a normal distribution, Mahalanobis distance performed much worse than psychometric antonyms and individual reliability. Meade and Craig (2012) warned that Mahalanobis distance may be sensitive to violations of multivariate normality, which could certainly occur when item response distributions are narrow or skewed.

Person-fit statistics based on item response theory (IRT) models have also been used to detect item score patterns that deviate significantly from expectations. In general, person-fit statistics aggregate differences between observed and expected item scores to determine whether a given respondent's pattern of item scores is consistent with the IRT model. The standardized log-likelihood l_z is commonly used to detect misfitting persons (Drasgow, Levine, & Williams, 1985). This statistic represents the likelihood of the observed item score pattern for a respondent with a certain estimated ability, transformed to a standard normal distribution, while accounting for differences in the variance of the likelihood distribution for respondents of different abilities. Such measures have been shown to have high detection rates for simulated random responding behavior on cognitive ability tests (Meijer, 2003), but research is lacking on the use of this approach for detecting IER on self-report inventories. One related study showed that the accuracy of IRT item parameter estimates was improved through an iterative process of removing respondents with poor person-fit when the data included simulated careless responders (Cheng, Patton, & Hong, 2018).

Critical values. A few IER indices have critical flagging values based on null-hypothesis significance testing. For example, the standardized log-likelihood l_z has a critical value of -1.65 below which a respondent is unlikely to have good person fit (one-sided test with Type-I error rate α = .05). Likewise, a respondent's squared Mahalanobis distance would be considered statistically significant if it exceeded a certain critical chi-squared value ($D^2 > \chi^2_{J,\alpha}$). In contrast, most IER-detection methods have empirically-derived cutoffs that can depend on the format of the instrument (e.g., number of response options, scale-point labels, number of items, item order, etc.) and the behavior of respondents in a particular sample. In the past, some researchers have skirted this issue by focusing on correlations among IER indices and other psychological variables, but cutoffs must be established to apply these methods in practice.

Cutoffs for long-string analysis may be based on the longest strings of respondents who are thought to be conscientious (Jackson, 1977), but in the absence of such data, Johnson (2005) proposed a "scree-like" approach to establishing cutoffs. This method involves identifying the longest continuous string of each response option for each respondent. The frequency distribution for each response option reveals the cutoff, which occurs at the last substantial decrease in the distribution before it becomes more uniform.

Specific cutoffs have been proposed for a few IER indices. Jackson (1977), for example, proposed that response patterns reflecting individual reliability less than .30 "can be categorized as probably primarily attributable to careless, non-purposeful, and/or inarticulated responding" (p. 41). This cutoff was based on the individual reliability distribution of simulated random responders, which had a mean of 0 and standard deviation of 0.18. Johnson (2005) applied a similar method to arrive at a cutoff of -0.03 for the psychometric antonyms index (with sign reversed). Specifically, he analyzed the psychometric antonym index distribution of "24,000 pseudo-random cases" and found that it had a mean of -0.02 and standard deviation of 0.18. Considering evidence that low-consistency response patterns may be valid, Johnson applied a conservative adjustment and determined that respondents with values less than -0.03 should be flagged. Another method—the response-operating curve (ROC) method—determines cutoffs by maximizing classification accuracy when detecting simulated random responding (Maniaci & Rogge, 2014). Note that cutoffs based on simulated random responding depend on the assumption that actual IER is well represented by random responding, which may not be defensible.

Although the proposed cutoffs for individual reliability and psychometric antonyms were determined using data from a particular instrument (Jackson, 1977; Johnson, 2005), the indices are based on correlations, which have the same scale regardless of the data source. Thus, those cutoffs might be applied to other instruments. Indeed, Huang and his colleagues (2012) used Jackson's (1977) and Johnson's (2005) cutoffs, but they also applied cutoffs that would identify 5% and 1% of simulated conscientious responders as exhibiting IER (like a Type-I error rate α = .05 or .01). This alternate approach mimics null-hypothesis significance testing using a simulated null distribution, and it depends on the assumption that simulated conscientious responding reflects the behavior of actual conscientious responders. Considering uncertainty about how IER manifests in real data, this assumption may be more tenable than Jackson (1977) and Johnson's (2005) assumption that IER can be simulated as random. Other researchers have applied normative cutoffs (e.g., 10% of respondents) in the absence of empirically justifiable ones (Dunn et al., 2016).

Effects of IER on Validity

IER invalidates the interpretation and use of scores for individuals, but the aggregate effects of IER on validity evidence for an instrument are uncertain. Random responding introduces measurement error that might be expected to attenuate correlations with other measures, reduce reliability, and distort factor structure (McGrath, Mitchell, & Kim, 2010). Error variance caused by IER could also reduce power in statistical analyses such as multiple regression (Maniaci & Rogge, 2014) and *t*-tests (DeSimone & Harms, 2017). However, IER is unlikely to be truly random, and nonrandom responding has unpredictable effects on validity evidence (Meade & Craig, 2012). Indeed, under certain circumstances, IER can even increase correlations among measures. For example, if inattentive responders complete two surveys (or subscale) such that IER leads to unusually low (or high) scores on both, the correlation between them would be inflated (Huang, Liu, & Bowling, 2015; Credé, 2010). IER tends to produce item scores closer to the center of the response scale (e.g., 3.5 on a 1–6 scale), so this is more likely to occur when item scores for typical conscientious responders are near the lower or higher end of the response scale.

In one prior study, coefficient alpha decreased slightly after removing respondents flagged by long-string and IRV methods (DeSimone & Harms, 2017). This result might be expected since choosing the same response to a large number of items reflects high internal consistency. Otherwise, prior research generally points to IER having small negative impacts on validity evidence. Other methods used in the DeSimone and Harms (2017) study increased alpha, and even filtering based on long-string and IRV methods increased correlations among subscales. In other studies, factor loadings for a personality inventory were lower for respondents with low individual consistency (Johnson, 2005), and correlations between personality factors were lower for respondents who missed attention-check items (Credé, 2010). Moreover, evidence of unidimensionality improved and coefficient alpha increased slightly after removing respondents flagged for IER (Huang et al., 2012), and coefficient alpha and multiple regression R^2 values were higher for attentive respondents (Maniaci & Rogge, 2014). In simulation research, random responding and socially-desirable faking had large effects on respondent’s scores and small negative effects on coefficient alpha and the correlation with a criterion measure (Zijlstra et al., 2011). Despite this evidence and perceptions that IER has moderate impacts on survey results (Liu, Bowling, Huang, & Kent, 2013), the use of IER detection methods is rare in published research (Ran, Liu, Marchiondo, & Huan, 2015).

As indicated by prior studies, removing data reflecting IER is generally expected to result in small changes to validity evidence. There are several possible explanations for this observation. First, respondents who exhibit IER may not do so consistently (Camus, 2015). If they are inattentive on a small percentage of items, their data would be difficult to distinguish from conscientious respondents. Even if such respondents are removed, the quality of their data would not have been bad enough to exert a discernable impact on validity evidence. Second, even if respondents who exhibit IER do so in extreme fashion (e.g., selecting the same response for every item), there may be so few such respondents that their impact on validity evidence would be minimal. Johnson (2005), for instance, flagged only 1% of respondents for having poor individual consistency and 0.9–3.5% for having long strings. The final explanation is that validity evidence may be quite strong, even ignoring the possible presence of IER. For example, there is little room for improvement by filtering out IER when coefficient alpha is .85 for a subscale and .93 for the entire test (DeSimone & Harms, 2017).

The Current Study

With increasing interest in measuring SEL competencies associated with college and career readiness comes increasing scrutiny of related assessments. This includes the basic question of whether students take such assessments seriously and if that is in doubt, whether individual results are trustworthy and whether provided validity evidence reflects an honest evaluation of an assessment’s utility. This study addressed those issues by attempting to identify students exhibiting IER and by estimating the effects of IER on validity evidence using a large data set compiled from numerous operational administrations of Engage, an SEL competency assessment taken by high school students. Results address the following research questions:

1. What is the base rate of IER in Engage data administered to high school students?

2. What effect does removing respondents exhibiting IER have on criterion-related validity coefficients, coefficient alpha, correlations among subscales, and confirmatory factor analysis model-data fit?

Results provide guidance for the use of IER detection in operational SEL assessments as well as methodological advice for future validity studies.

Method

Measure

Engage (ACT, 2016) is a psychological inventory including 108 items divided among 10 subscales relating to three broad domains: motivation, self-regulation, and social engagement (Table 1). Responses reflect a 6-point Likert scale from Strongly Disagree to Strongly Agree, and 31 items have negative wording. Engage was identified by Herman and Hilton (2017) as one of two SEL assessments with “extensive validation research” (p. 97). The factor structure with 10 first-order factors and three second-order factors (Table 1) was identified and validated by Le, Casillas, Robbins, and Langley (2005). A subsequent study revealed that academic discipline, social activity, and steadiness provided incremental improvement to the prediction of first-year college grade-point average and persistence to the second year of college, over that provided by prior academic achievement (Robbins, Allen, Casillas, Peterson, & Le, 2006).

Sample Description

The data analyzed in this study comprised 18,578 records from high school students who took Engage sometime between 2009 and 2013. The sample was 52% female, 53% White, 16% Black, 16% Hispanic, and 4% Asian. By grade level, the sample included 27% 10th graders, 44% 11th graders, and 19% 12th graders. When students took Engage, 96% of them self -reported their cumulative high school grade point averages (HSGPA). The distribution of HSGPA was 28% A- to A, 25% B to B+, 16% B- to B, and 15% C to B-, which reflected slightly lower HSGPAs than the national sample of high school graduates who took the ACT college admissions tests (33%, 24%, 15%, and 9%, respectively). This was expected since the sample included students who did not eventually graduate high school or take the ACT. ACT scores were available for 15% of the sample. The average ACT Composite score was 21.2, which was close to the national average of 20.8.

Table 1. Engage Scales

Domain	Scale	No. Items	Score Range	Definition
Motivation	Academic Discipline	10	10–60	The amount of effort a student puts into schoolwork and the degree to which he or she sees himself or herself as hardworking and conscientious
	Commitment to College	10	10–60	Commitment to staying in college and getting a degree
	Communication Skills	10	10–60	Attentiveness to others’ feelings and flexibility in resolving conflicts with others
	General Determination	11	11–66	The extent to which a student strives to follow through on commitments and obligations
	Goal Striving	10	10–60	The strength of one’s effort to achieve objectives and end goals
	Study Skills	12	12–72	The extent to which a student believes he or she knows how to assess an academic problem, organize a solution, and successfully complete academic assignments
Social Engagement	Social Activity	10	10–60	One’s comfort in meeting and interacting with other people
	Social Connection	11	11–66	One’s feelings of connection and involvement with the college or school community
Self-Regulation	Academic Self-confidence	12	12–72	The belief in one’s ability to perform well in school
	Steadiness	12	12–72	One’s response to and management of strong feelings

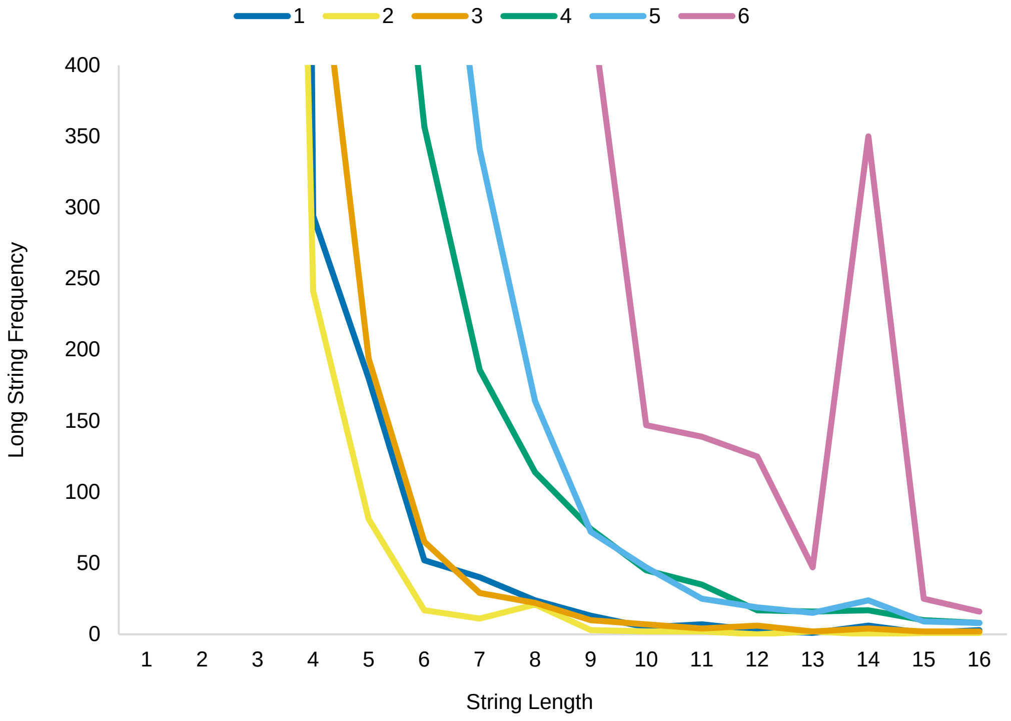
Analysis

IER detection methods. A total of nine IER indices were calculated using Engage data: mean absolute difference between positively- and negatively-worded items (MAD), maximum item score frequency, inter-item standard deviation (ISD), intra-individual response variability (IRV), long-string (LS), resampled individual reliability (RIR), psychometric antonyms (PA), squared Mahalanobis distance (*D*^{2}), and the standardized log-likelihood (l_z). Note that RIR was calculated as the mean of Spearman-Brown adjusted correlations for 30 random splits of items in the 10 Engage subscales. The sign of the PA was reversed so that respondents with low PA indices would be suspected of IER. Considering the factor structure of Engage (Table 1; Le et al., 2005), l_z was based on a multidimensional graded response model (Samejima, 1969) with three latent traits (motivation, social engagement, and self-regulation). Each method was applied individually to flag respondents exhibiting IER. Then, since different methods are better or worse at detecting certain types of IER (Meade & Craig, 2012), several approaches were applied simultaneously.

IER index cutoffs. For the standardized log-likelihood and Mahalanobis distance, respondents were flagged using cutoffs based on null-hypothesis significance testing (with Type-I error rate $\alpha = .01$). Johnson's (2005) "scree-like" approach was used to determine cutoffs for long-string analysis, which were 5, 5, 6, 9, 10, and 14 for response options 1–6, respectively (Figure 1). The high cutoff for response option 6 was due to a large number of respondents choosing the sixth option on 14 consecutive positively-worded items.

The Engage cutoff of 90% was applied for maximum item score frequency. For other IER indices, the proposed method of simulating the null distribution of an IER index was used to determine cutoffs (Huang et al., 2012). First, conservative data cleaning procedures were applied to remove respondents with more than 30% missing data or who were in the 3% most likely exhibiting IER according to each detection method (typical IER rates are slightly higher; Meade & Craig, 2012). In this process, the long string cutoffs were increased by 2 to make the flagging criteria more conservative. Consistent with the method of calculating l_z, a three-dimensional graded response model was fit to the remaining data. The estimated item parameters were then used to simulate item scores for 10,000 simulated respondents with ability parameters drawn from the distribution of actual respondents. The distributions of IER indices based on the simulated data revealed the cutoffs corresponding to a 1% Type-I error rate ($\alpha = .01$).

Figure 1. "Scree-like" plot of long string frequencies



Estimating the effects of IER. To estimate the prevalence of IER, the percentage of respondents flagged by each detection method was calculated. Agreement between different methods was evaluated using correlations, bivariate scatter plots, and classification consistency. Then, to estimate the effects of IER on validity, four types of validity evidence were generated: criterion-related validity coefficients, coefficient alpha, correlations among subscales, and confirmatory factor analysis fit indices (assuming 10 first-order factors and three second-order factors; Le et al., 2005). These measures were each calculated three times: using all available data, using only respondents flagged for IER, and using only respondents not flagged for IER. Correlations calculated on subsamples of respondents were adjusted for possible restriction of range. The average of coefficient alpha for the 10 Engage subscales was calculated. Similarly, the average correlation among the Engage subscales (adjusted for restriction of range) served as an indicator of convergent validity. Finally, confirmatory factor analysis (CFA) was applied to the data using full information maximum likelihood estimation implemented with the R package *lavaan* (Rosseel, 2012), assuming the hierarchical factor structure identified by Le and his colleagues (2005). To support model convergence, the items in each subscale were divided into three item parcels. Models were compared using the following model-data fit indicators: root mean square error of approximation (Steiger & Lind, 1980), comparative fit index (Bentler, 1990), and the standardized root mean square residual (Hu & Bentler, 1999).

Results

Engage Descriptive Statistics

Table 2 shows descriptive statistics for Engage scores, correlations between Engage scores and four criterion variables, and coefficient alpha. The average item score was 4.6, which is reflected in the generally high scores on the various Engage scales. The highest mean item scores were observed for General Determination and Commitment to College, and the lowest were observed for Steadiness and Social Activity. With a mean of .33 across the 10 subscales, Engage scores correlated highest with HSGPA. The mean correlations were .18, .15, and .12 for homework, absence, and ACT Composite, respectively. Subscale coefficient alphas ranged from .83 to .89 with a mean of .86. Table 3 shows the correlations among Engage subscales. Goal Striving had the highest average correlation with the other subscales (.58); Social Activity had the lowest (.32). The average correlation among Engage subscales within domains (motivation, social engagement, and self-regulation) was .59 whereas the average correlation across domains was .39.

Table 2. Engage Subscale Descriptive Statistics, Correlations with Criterion Variables, and Reliability

				Correlation with Criterion*				
Scale	Mean	SD	Mean Item Score	HSGPA	Homework	Absence	ACT	Coef. alpha
Academic Discipline	46.2	9.2	4.6	.56	.32	.26	.13	.87
Commitment to College	51.2	8.9	5.1	.40	.20	.17	.09	.88
Communication Skills	48.9	7.6	4.9	.26	.16	.14	.10	.85
General Determination	56.7	7.5	5.2	.34	.24	.15	.04	.88
Goal Striving	49.9	7.7	5.0	.33	.22	.15	.04	.87
Study Skills	51.8	10.9	4.3	.25	.18	.12	.05	.89
Social Activity	40.8	9.8	4.1	.15	.07	.08	.13	.85
Social Connection	48.6	9.7	4.4	.29	.12	.17	.08	.83
Academic Self-confidence	51.7	10.5	4.3	.48	.18	.15	.45	.84
Steadiness	48.1	11.8	4.0	.23	.15	.15	.14	.87
Motivation	304.6	42.5	4.8	.44	.27	.20	.09	.96
Social Engagement	89.4	16.9	4.3	.25	.11	.14	.12	.88
Self-Regulation	99.8	19.0	4.2	.41	.19	.17	.34	.89
Total	493.8	66.6	4.6	.46	.25	.21	.19	.97

*** The correlations sample sizes were approximately 17,800 (HSGPA), 14,800 (Homework), 13,600 (Absence), and 2,800 (ACT).**

Table 3. Correlations Among Engage Subscales

Scale	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Academic Discipline										
2. Commitment to College	.59									
3. Communication Skills	.49	.46								
4. General Determination	.71	.58	.68							
5. Goal Striving	.66	.59	.61	.82						
6. Study Skills	.57	.45	.62	.66	.68					
7. Social Activity	.22	.25	.31	.28	.38	.19				
8. Social Connection	.43	.41	.60	.50	.55	.44	.52			
9. Academic Self-confidence	.52	.42	.31	.46	.54	.40	.38	.30		
10. Steadiness	.43	.33	.40	.38	.41	.39	.35	.26	.45	
Mean	.51	.45	.50	.56	.58	.49	.32	.45	.42	.38

Note: All correlations were significant at the *p* < .001 level.

IER Index Descriptive Statistics

Univariate distributions. Table 4 shows descriptive statistics for the distributions of the IER indices. The mean absolute difference between positively- and negatively-worded items was 0.75. The positively-worded mean was greater than the negatively-worded mean for 88% of respondents, which could indicate acquiescence bias toward the "agree" end of the response scale. On average, respondents had the same item score on 45 of the 108 items. ISD and IRV were similarly distributed except that IRV values were slightly higher because they were calculated before rescoring the negatively-worded items. On average, respondents' longest strings of the same response were 6.33 consecutive items.

Most respondents had high, positive RIR, suggesting consistent responding behavior within subscales. For many respondents, the individual reliability index varied considerably across the 30 iterations used to calculate RIR (the average standard deviation was 0.17). Thus, concerns about the instability of individual reliability estimates based on a single splitting of the subscale items (Curran, 2016) seem to be warranted. In general, item scores on the 30 psychometric antonym item pairs correlated as expected for conscientious respondents (i.e., negatively, which resulted in a positive mean since the sign was reversed). The distribution of D^2 indicates that most respondents' item scores deviated from the average item scores but not to a statistically significant extent. Finally, the average l_z value was close to the expected value of 0, and most values were greater than the critical value of -2.33.

Bivariate distributions. Figure 2 illustrates the associations among the IER indices with scatter plots and correlations. Spearman correlations were calculated since most of the relationships were nonlinear. Note that strong or weak correlations can indicate that the indices detect similar or different types of IER, respectively.

Table 4. Descriptive Statistics for IER Index Distributions

Index	Mean	SD	25th Percentile	Median	75th Percentile
MAD	0.75	0.59	0.31	0.63	1.04
Max. Pct.	45.10	14.87	34.30	42.30	53.70
ISD	1.29	0.31	1.07	1.27	1.49
IRV	1.59	0.32	1.39	1.61	1.81
LS*	6.33	7.33	4.00	5.00	7.00
RIR	0.63	0.25	0.55	0.70	0.80
PA	0.69	0.27	0.56	0.77	0.90
D^2	108.15	64.77	64.49	92.24	132.69
l_z	0.19	2.39	-0.99	0.57	1.78

Note: MAD = mean absolute difference, Max. Pct. = maximum percentage of the same item score, ISD = inter-item standard deviation, IRV = intra-individual response variability, LS = long string, RIR = resampled individual reliability, PA = psychometric antonyms, D^2 = Mahalanobis distance, and l_z = standardized log-likelihood.
*** There is no single long string index, so the distribution of the respondents' longest strings is reported here.**

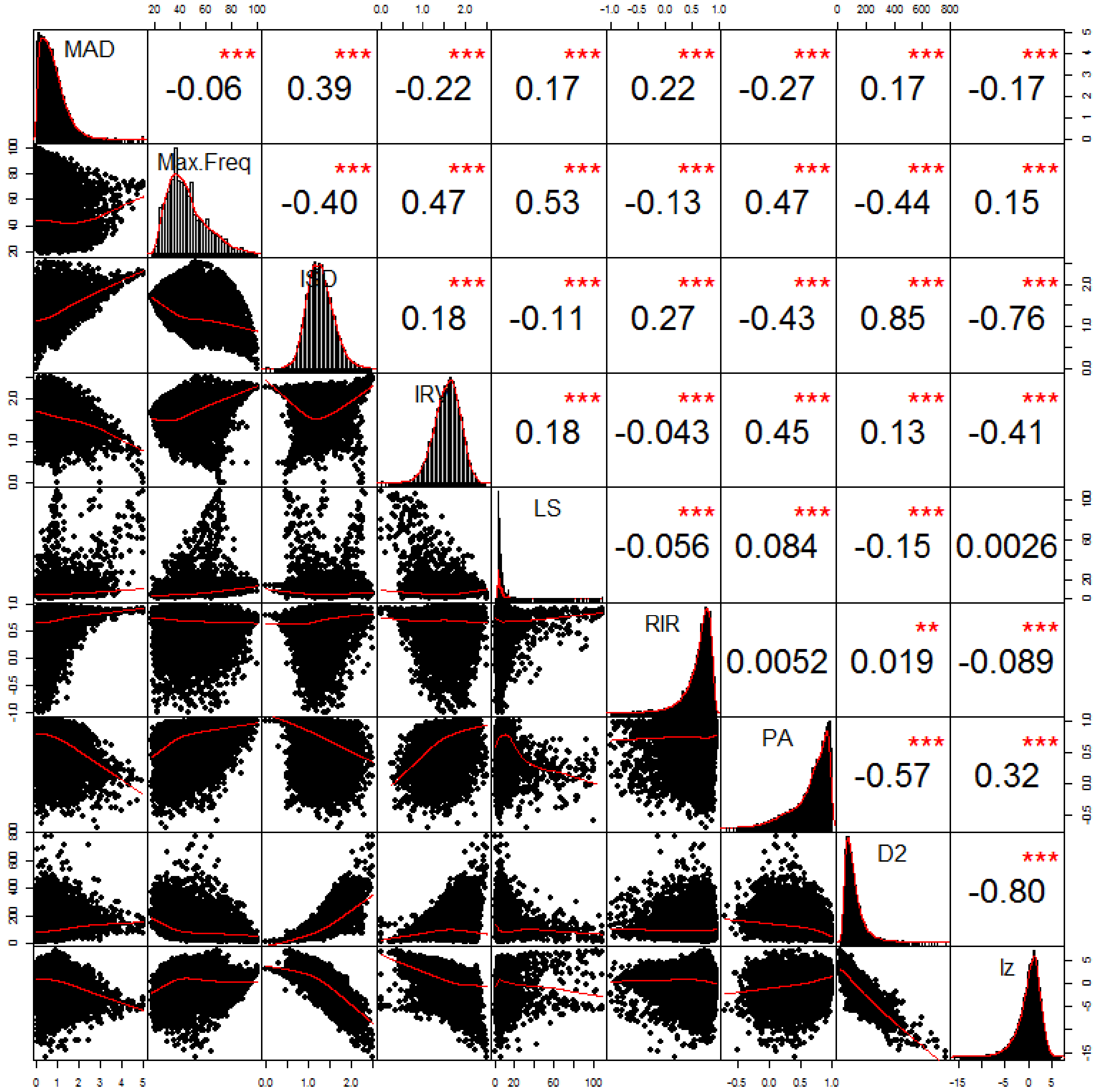
The correlations generally indicated that evidence of IER from one index (e.g., high MAD) was associated with evidence of IER from another index (e.g., low PA index). ISD, D^2, and l_z were most strongly related, with correlations among them of .85, -.76, and -.80. Specifically, respondents with high ISD tended to have high D^2 and low l_z. These would be respondents with widely varying item scores, which apparently diverged from the mean item score vector and differed significantly from expectations based on an IRT model.

The correlations indicated that respondents with high maximum frequency tended to have low ISD, low IRV, and long strings, which would be expected since high maximum frequency indicates low response variability. However, those respondents also tended to have high PA indices and low D^2, which do not suggest IER. The scatter plots indicated that low ISD and high ISD may detect different sorts of IER. Specifically, high ISD was associated with higher MAD (.39), lower PA index (-.43), higher D^2 (.85), and lower l_z (-.76), but low ISD was associated with higher maximum frequency (-.40). MAD correlated moderately with ISD (.39), and IRV correlated moderately with the maximum frequency, PA index, and l_z (.47, .45, and -.41, respectively). Respondents' longest strings correlated weakly or negligibly with indices other than maximum frequency (.53), but a long string is not necessarily indicative of IER; it depends on which response option the long string occurred. The PA index correlated moderately with the other indices except LS and RIR.

One unexpected result was apparent from Figure 2: RIR correlated positively with MAD (.22) and ISD (.27), and it correlated negligibly with all other IER indices. For reference, individual reliability correlated .35 with PA and -.57 with D^2 in a prior study (Meade & Craig, 2012). In theory, respondents with high RIR are conscientious, and they should not have high MAD or high ISD. Yet, an actual respondent who selected 1 for every item (MAD = 5, ISD = 2.27) had RIR of .94. Further investigation revealed a unique feature of Engage that caused this result. Specifically, the negatively-worded items were not evenly distributed across the subscales.

Indeed, the 31 negatively-worded items were heavily concentrated within three of the subscales. Thus, even though the example respondent had inconsistent item scores across the assessment (1s and 6s after negative scoring), he or she had highly consistent item scores within most subscales. An examination of the scatter plots in Figure 2 revealed that respondents with low RIR tended to have low MAD, moderate ISD, moderate IRV, short strings, high PA index, low D^2, and high l_z. In other words, low RIR appeared to be a better indicator of conscientious responding than IER.

Figure 2. Histograms, correlations, and scatter plots for IER indices.



Characteristics of respondents likely exhibiting IER. For each IER index, demographic characteristics of the 5% of respondents most likely exhibiting IER were compared to the total sample (Table 5). This analysis would indicate how filtering for IER might affect sample representation. There was a common pattern observed for MAD, high ISD, IRV, LS, PA, D^2, and l_z. Namely, flagged respondents were more likely to be male, less likely to be White, less likely to speak English at home, and they tended to have lower grades in high school, lower average ACT Composite scores, and lower Engage scores. Compared to Hispanic respondents, African American respondents were more likely to be flagged for MAD, high ISD, D^2, and l_z, but less likely to be flagged by IRV and PA. In complete contrast, respondents flagged for maximum item score frequency and low ISD were more likely to be female and White and they tended to earn higher grades, ACT scores, and Engage scores. No noteworthy demographic differences were apparent for respondents flagged for RIR.

Table 5. Percent Difference or Mean Difference between Most Likely IER Respondents and Full Sample

Index	Female	African American	Hispanic	White	English	A- to A HSGPA	ACT Composite	Engage Total
MAD	-5.6%	8.9%	4.8%	-16.6%	-6.4%	-17.2%	-3.9	-21
Max. Freq.	5.9%	5.7%	-4.8%	3.3%	1.3%	19.8%	0.9	112
ISD (low)	3.4%	-5.9%	-3.3%	11.2%	3.8%	25.5%	2.8	92
ISD (high)	-4.5%	9.3%	-1.3%	-10.4%	-1.4%	-16.1%	-0.7	-61
IRV	-19.3%	-3.2%	10.1%	-12.8%	-9.1%	-20.3%	-2.1	-80
RIR	-0.8%	-2.0%	0.4%	2.9%	-2.4%	-2.1%	0.2	13
LS	-16.4%	3.4%	2.5%	-10.1%	-2.5%	-16.6%	-1.8	-70
PA	-15.7%	-0.3%	7.1%	-14.2%	-10.9%	-23.5%	-1.7	-91
D^2	-12.9%	7.0%	1.2%	-13.0%	-5.4%	-18.7%	-0.7	-89
l_z	-15.1%	8.6%	-0.6%	-12.5%	-5.2%	-17.6%	-0.5	-69

Selecting methods. Considering results from the descriptive analyses, three IER indices were excluded from the remainder of the study. RIR was excluded because it did not apparently indicate IER for Engage data; maximum item score frequency and low ISD were excluded because of their tendencies to remove high achieving students, many of whom also had high Engage scores. Even though some respondents flagged by the latter two methods might have exhibited IER (e.g., socially desirable responding), the flagged respondents apparently included conscientious respondents with high academic achievement and high SEL competencies. Moreover, only 129 respondents (0.7%) had a maximum item score frequency of 90% or more, so their effect on validity coefficients would have been negligible.

IER Prevalence

Flagging cutoffs were estimated from simulated response data based on a three-dimensional graded response model fit using a cleaned data set. The cleaning process removed 13.7% of the data due to suspected IER. Table 6 shows the $\alpha = .01$ cutoffs for the IER indices and the resulting percentages of flagged respondents. PA flagged the fewest respondents (0.8%), and the PA cutoff was low compared to the cutoff used in prior research (-.19 vs. -.03). On the other hand, some IER indices flagged a large number of respondents. Using a cutoff of 1.29, 16.7% of respondents were flagged for low IRV. More than 20% were flagged for having a statistically significant D^2, but the chi-squared statistical test for D^2 was likely very sensitive on account of the large number of items. In all, 42.8% of respondents were flagged by at least one of the IER detection methods.

Table 6 also provides example flagged item score patterns, which reveal some commonly flagged patterns. For instance, respondents who ignored negatively-worded items and selected the same response option very frequently could have been flagged by any of the IER indices. IRV also flagged respondents with consistently mid-range responses (i.e., 3s and 4s). Methods such as ISD, D^2, and l_z detected unexpectedly high variability (e.g., a broad mix of item scores after rescoring).

Table 6. IER Prevalence and Example Flagged Item Score Patterns

Index	Cutoff	% Flagged	Example Flagged Item Score Patterns (Items 1–37)
MAD	≥ 1.49	10.1%	1111111111111111111111111111111111111 (raw) 6463664143616415326454513656464546564 (rescored)
ISD	≥ 1.63	14.0%	1161161666636116116666616666111666611 (rescored) 5362155566436365444125611665151144544 (rescored)
IRV	≤ 1.29	16.7%	66655556666255555665 6666655554455666 (raw) 4355345333443433334433432443434434343 (raw)
LS		5.0%	5164414444444444444444444665444452354 (raw) 6 666666646665665666666666666666666666 (raw)
PA	< -0.19	0.8%	3163234223323323444444444555555555555 (raw) 655545554 5454345566565543345545556 (raw)
D^2	> 145.10	20.3%	6666664666356663666333666666636626666 (rescored) 5561332424314351256344512656454436465 (rescored)
l_z	< -2.33	13.3%	5362155566436365444125611665151144544 (rescored) 2453245313315333354531544655336516414 (rescored)

Flagging agreement among methods was examined by calculating the percentage of respondents flagged by one method also flagged by another method. Table 7 shows, for example, that of the respondents flagged for MAD, 44% were also flagged for high ISD. The greatest agreements were among high ISD, D^2, and l_z, which should be expected given the correlations in Figure 2. LS and IRV are intended to detect similar sorts of IER, and agreement between them was apparent in Table 6. Agreement for PA was difficult to evaluate since so few respondents were flagged by that method, but PA flagging apparently overlapped with all other methods. IRV had the least overlap with other methods.

Table 7. Percentage Agreement between IER Flags*

Index	MAD	ISD (high)	IRV	LS	PA	D^2	l_z
MAD	--	44%	28%	15%	3%	33%	31%
ISD (high)	32%	--	5%	13%	2%	80%	64%
IRV	17%	4%	--	15%	3%	6%	5%
LS	31%	36%	51%	--	4%	37%	38%
PA	32%	28%	56%	26%	--	51%	42%
D^2	17%	55%	5%	9%	2%	--	57%
l_z	23%	68%	6%	14%	3%	87%	--

*** For example, of the respondents flagged for MAD, 44% were also flagged for high ISD, 28% were also flagged for low IRV, etc.**

Changes in Validity Evidence

Since PA flagged only 0.8% of respondents, removing those respondents was not expected to affect validity evidence. To detect a possible effect for PA, a normative cutoff that flagged 5% of respondents was applied in the validity evidence analysis (PA index < -.12).

Validity coefficients. In the analysis of validity coefficients, the mean correlation between the 10 Engage subscales and each of the four criterion variables was calculated. The "IER only" columns of Table 8 show the differences in average correlations between flagged respondents and all respondents. These values are negative for all IER detection methods, which indicates that validity coefficients were lower for the flagged respondents. Differences between flagged and unflagged respondents manifested in the very small differences shown in the "IER Removed" columns of Table 8. That is, removing respondents suspected of IER had little effect on the validity coefficients. The average correlation increased by .03 at most, and typical changes were smaller.

Table 8. Mean Differences in Validity Coefficients Across 10 Engage Scales Relative to Baseline*

		IER Only				IER Removed			
Index	Cutoff	HSGPA	HW	Abs.	ACT	HSGPA	HW	Abs.	ACT
MAD	≥ 1.49	-.15	-.07	-.05	-.13	.02	.01	.01	.01
ISD	≥ 1.63	-.21	-.07	-.08	-.16	.03	.01	.01	.02
IRV	≤ 1.29	-.08	-.06	-.04	-.14	-.01	-.01	.00	-.01
LS		-.14	-.07	-.02	-.17	.00	.00	-.01	.01
PA	< -0.12**	-.23	-.08	-.08	-.30	.00	.00	.00	.01
D^2	> 145.10	-.17	-.06	-.06	-.16	.02	.01	-.01	.03
l_z	< -2.33	-.21	-.09	-.07	-.15	.03	.02	.00	.03

Coefficient alpha. The average coefficient alpha of the 10 Engage subscales was calculated to examine possible changes in alpha from removing suspected IER. Table 9 shows that, for all methods, coefficient alpha was lower for the flagged respondents. Respondents flagged for PA had the lowest average coefficient alpha, which might be expected since PA should flag respondents with inconsistent scores on pairs of positively- and negatively-worded items (after negative scoring). Though the flagged respondents exhibited lower internal consistency, removing them had negligible effects on the average coefficient alpha of the subscales. This result might have been expected since there was little room for improvement in the coefficient alphas, which ranged from .83 to .89.

Subscale correlations. For all IER indices, the average correlation among the 10 Engage subscales was lower for flagged respondents than for the full sample, but removing them had little effect on the average correlations (Table 9). The biggest changes were observed for high ISD, D^2, and l_z (.04–.06), which were the three indices that flagged the largest percentages of respondents.

Table 9. Differences in Validity Evidence Relative to Baseline[*]

		Mean Coefficient Alpha Difference		Mean Subscale Correlation Difference		CFA Model-Data Fit Difference (IER Removed)		
Index	Cutoff	IER Only	IER Removed	IER Only	IER Removed	RMSEA	CFI	SRMR
MAD	≥ 1.49	-.06	.00	-.13	.03	-0.01	0.02	-0.01
ISD	≥ 1.63	-.04	.00	-.27	.06	-0.01	0.02	-0.01
IRV	≤ 1.29	-.05	-.01	-.15	-.01	-0.01	0.01	0.00
LS		-.02	-.01	-.09	.01	0.00	0.01	-0.01
PA	< -0.12[**]	-.19	.00	-.22	.00	0.00	0.00	0.00
D^2	> 145.10	-.05	.01	-.18	.04	0.00	0.00	0.00
l_z	< -2.33	-.05	.01	-.23	.05	0.00	0.01	-0.01

Factor analysis model-data fit. Engage factor model convergence was poor for the flagged respondents alone, which was likely due to smaller sample size and response behavior inconsistent with the model. Table 9 shows differences in model-data fit indices before and after filtering the data. In general, the non-zero changes suggest very small improvements to model-data fit by removing suspected IER (i.e., RMSEA decreased, CFI increased, and SRMR decreased). A 0.01 change in CFI is considered a meaningful difference (Cheung & Rensvold, 2002), so filtering with MAD, ISD, IRV, LS, and l_z all seemed to effect some improvement in model-data fit.

Combining methods. IER detection methods were applied simultaneously in an attempt to flag several types of IER. One approach would involve removing respondents flagged by two or more methods (21.2%) or those flagged by three or more methods (11.0%). Improvements to validity evidence were similar for both options, so results for three or more are reported here.

As in the preceding analyses, the flagged respondents had lower average criterion-related validity coefficients (.08 to .22 lower), a lower average coefficient alpha (.06 lower), and a lower average correlation among subscales (.26 lower). The effects of removing the flagged respondents on validity evidence were still small. Average correlations with external criteria changed by .00–.03, the average coefficient alpha was unchanged, the average subscale correlation increased by .04, and CFA model-data fit improved slightly (e.g., CFI increased by .01). Those changes were similar in magnitude to individual detection methods, but only 11.0% of respondents were removed, which was less than several individual methods.

A different approach would remove respondents flagged by any of the methods. To avoid flagging an impractically high percentage of respondents (i.e., 42.8%), normative cutoffs that would flag 5% were applied to all methods except LS. This resulted in 14.9% of respondents being flagged. When that group of respondents was removed, average criterion-related validity coefficients increased by .01–.02, coefficient alpha increased by an average of .01, the average correlation among subscales increased by .02, and CFI increased by .01.

Omitted methods. As an exploratory exercise, changes to validity evidence were evaluated for the omitted IER detection methods. The effects of filtering based on maximum item score percentage and RIR were very small and not in consistent directions. For low ISD, however, the validity coefficients were .19–.46 higher for respondents in the bottom 5% of the ISD distribution. This result might be expected considering that respondents with low ISD tended to have high Engage scores, high grades, and high ACT scores. Respondents flagged for low ISD also had an average coefficient alpha that was .09 higher than the full sample, which is explained by the fact that low ISD reflects high internal consistency (i.e., the same or similar score on all items). Subscale correlations were .31–.36 higher for respondents flagged for low ISD. Again, these respondents tended to have consistent scores across items and subscales, which would manifest in high correlations. Finally, confirmatory factor analysis model-data fit was slightly worse after removing respondents flagged for low ISD, which was not surprising given that uniform item scores would be consistent with the factor model.

Discussion and Conclusions

The initial analyses examined the distributions of the IER indices, their bivariate relationships, and the demographic characteristics of respondents most likely to exhibit IER. Some methods correlated strongly (e.g., high ISD, D^2, and l_z), which is consistent with the notion that those methods capture similar sorts of IER. Results also revealed an important lesson about the application of IER detection methods: no method is guaranteed to work for a given assessment or testing population. In the case of RIR, a design feature of Engage caused RIR to produce nonsensical results (i.e., very high RIR for individuals obviously lacking in internal consistency). The same problem with RIR is likely to occur whenever the subscales of an assessment are each comprised primarily of positively-worded or negatively-worded items.

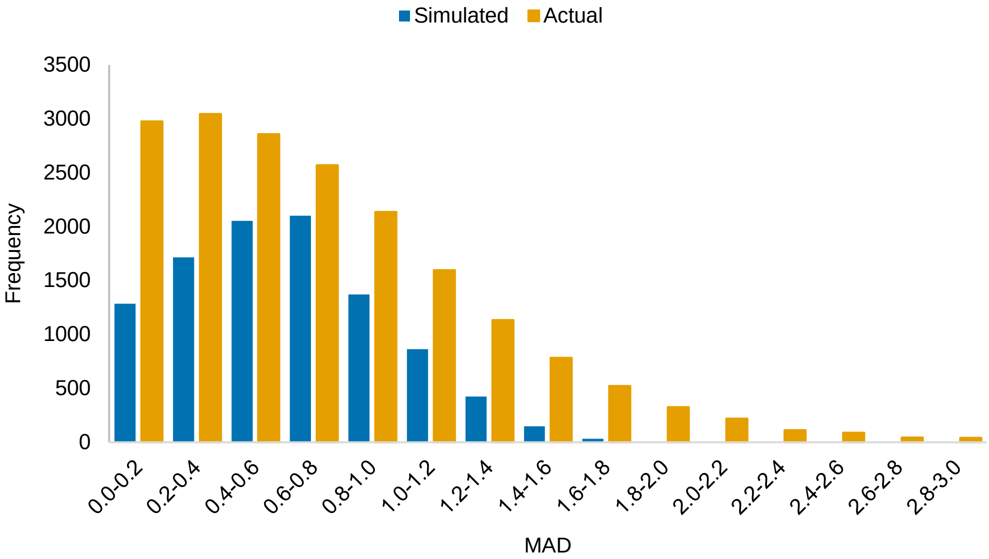
Another difficulty arose with maximum item score frequency and low ISD, both of which indicated invariant item scores. Respondents with low item score variability often had very high Engage scores, which could be indicative of socially-desirable responding. However, many such respondents also had high grades and high ACT scores, which makes them plausibly conscientious. Thus, it would be inappropriate to flag and remove their data. Indeed, subsequent analyses revealed that removing their data caused validity evidence to weaken slightly. In future validity studies, researchers should be aware that invariant responding is likely to increase some validity indicators. Additional analyses should determine whether that increase is primarily legitimate (due to conscientious responding) or spurious (due to IER).

The first research question concerned the prevalence of IER among high school students on a measure of SEL competencies related to college and career readiness. Estimates ranged from 0.8% to 20.3% depending on which IER detection method was applied. Combined, this resulted in 42.8% of all respondents being flagged by one or more methods. Estimates from most prior research were approximately 10% or lower (Meade & Craig, 2012; Maniaci & Rogge, 2014), but values as high as 50% have been reported (Curran, Kotrba, & Denison, 2010). Naturally, prevalence should be expected to differ across assessments and samples, but prevalence estimates also depend on decisions and assumptions directly affecting the flagging cutoff values. For this study, simultaneous consideration of improvements to validity evidence and minimization of the number of respondents removed could guide the examination of IER prevalence estimates. In that line of reasoning, the best estimate of IER prevalence from this study was 11.0%, which was achieved by removing respondents flagged by three or more detection methods. This approach achieved improvements to validity evidence at least as large as individual detection methods while removing fewer respondents. Moreover, using multiple flags has the benefit of increasing confidence in the decision to remove respondents.

In this study, the cutoff criteria for long string analysis were based on the author’s reading of frequency distributions. For other methods, a Type-I error rate had to be selected. A conservative rate of 1% was chosen to minimize false positive flags, yet some detection methods still flagged a high percentage of students. This was especially true of Mahalanobis distance, which flagged 20.3% of respondents. However, the chi-squared significance test for Mahalanobis distance was particularly sensitive on account of the large number of items. Thus, an even more conservative Type-I error rate might have been appropriate to counteract this other factor influencing estimated prevalence.

In this study and others (Huang et al., 2012), some of the flagging cutoff values were based on simulated conscientious respondents. This procedure depends on the selection of an IRT model and the assumption that simulated respondents behave similarly to actual respondents. As an example, consider the distributions of MAD shown in Figure 3. Notice that the simulated distribution had very few respondents with MAD values greater than 1.5, which would be appropriate for conscientious respondents. However, compared to the actual distribution, the simulated distribution had a smaller proportion with MAD values in the 0.0–0.5 range. Thus, the simulated distribution was lacking in a certain type of conscientious respondent: those scoring very consistently across positively- and negatively-worded items. Unfortunately, without knowledge of the true distribution for conscientious respondents, it is difficult to ascertain when a violation of assumptions results in a Type-I error rate other than intended.

Figure 3. Histograms of simulated and actual distributions of the mean absolute difference (MAD) index.



The second research question focused on the effects of IER on Engage validity evidence. MAD, high ISD, IRV, LS, PA, D^2, and l_z each identified respondents whose data reflected lower criterion-related validity coefficients, lower coefficient alpha, and lower convergent validity coefficients. Removing those respondents from the data had the effect of improving those three types of validity evidence as well as confirmatory factor analysis model-data fit. Consistent with prior research, these effects were quite small (Huang et al., 2012; Maniaci & Rogge, 2014; Zijlstra et al., 2011). This finding is likely caused by a combination of factors: low IER prevalence, low IER severity, and strong validity evidence even with IER present. In future validity studies, researchers should consider presenting results before and after removing suspected IER. They must also consider whether the possible benefits of removing apparent IER outweigh the associated loss in statistical power in subsequent analyses, especially when overall sample sizes are small.

Even if IER has little impact on validity evidence, IER detection methods can still reasonably be applied to flag individual results when administering SEL assessments operationally. The seven methods investigated here all apparently identified data that included IER. As long as self-report SEL assessments are administered under low-stakes conditions, there is no need to “invalidate” results for flagged respondents. However, it may be helpful to instill a healthy degree of skepticism when interpreting certain results, especially those flagged by multiple IER detection methods. In any case, test administrators should cross-check results with other indicators of SEL competencies. For example, very high Engage scores for a student with very low grades might be suspicious.

In this study, IER detection methods were applied in a new and important context: self-report assessments of SEL competencies related to college readiness. This study introduced the three Engage methods (MAD, maximum item score frequency, and low ISD) to the research literature and evaluated their use when reporting Engage results. Analyses illustrated that IER detection methods cannot be assumed to work as advertised, and their use may have unexpected, negative consequences. This study generated estimates of IER prevalence similar to those from other contexts, and it illustrated the difficulties inherent in estimating IER prevalence. Moreover, results corroborated prior studies showing that the effects of IER on validity evidence tend to be quite small. Even so, researchers and test administrators must be attentive to IER because its effects could be greater for other assessments. Moreover, operational assessment programs can still use IER detection methods to flag individual results as potentially invalid.

References

ACT. (2016). *Development and validation of ACT Engage: Technical manual.* Iowa City, IA: ACT. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/act-engage-technical-manual.pdf

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Austin, E. J., Deary, I. J., Gibson, G. J., McGregor, M. J., & Dent, J. B. (1998). Individual response spread in self-reported scales: Personality correlations and consequences. *Personality and Individual Differences, 24*(3), 421–438. doi:10.1016/S0191-8869(97)00175-X

Baumeister, R. F., & Tice, D. M. (1988). Metatraits. *Journal of Personlity, 56*(3), 571–598. doi:10.1111/j.1467-6494.1988.tb00903.x

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Camara, W., O'Connor, R., Mattern, K., & Hansen, M. A. (2015). *Beyond academics: A holistic framework for enhancing education and workplace success.* (ACT Research Report Series 2015-4). Iowa City, IA: ACT. Retrieved from http://www.act.org/content/dam/act/unsecured/documents/ACT_RR2015-4.pdf

Camus, K. A. (2015). *Once careless, always careless? Temporal and situational stability of insufficient effort responding (IER).* (Unpublished master's thesis). Dayton, OH: Wright State University. Retrieved from http://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=2757&context=etd_all

Cheng, Y., Patton, J., & Hong, M. (2018). *Detection and treatment of careless responses to improve item parameter estimation.* Notre Dame, IN: University of Notre Dame. Manuscript under review.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluting goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Conley, D. T. (2007). *Redefining college readiness.* Eugene, OR: Education Policy Improvement Center. Retrieved from http://files.eric.ed.gov/fulltext/ED539251.pdf

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment: Personality measurement and testing* (pp. 179–198). London: SAGE.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*(4), 596–612. doi:10.1177/0013164410366686

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19. doi:10.1016/j.jesp.2015.07.006.

Curran, P. G., Kotrba, L., & Denison, D. (2010, April). *Careless responding in surveys: Applying traditional techniques to organizational settings.* Paper presented at the 25th annual conference of the Society for Industrial/Organizational Psychology, Atlanta, GA.

DeSimone, J. A., & Harms, P. D. (2017). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, 1–19. doi:0.1007/s10869-017-9514-9

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(1), 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. doi:10.3102/0013189X15584327

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2016). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*(1), 1–17. doi:10.1007/s10869-016-9479-0

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*(1), 82–98.

Herman, J., & Hilton, M. (2017). *Supporting students' college success: The role of assessment of interpersonal and interpersonal competencies.* Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/24697

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi:10.1080/10705519909540118

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*(2), 299–311. doi:10.1007/s10869-014-9357-6

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*(1), 99–114. doi:10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828–845. doi:10.1037/a0038510

Jackson, D. N. (1976). *The appraisal of personal reliability.* Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey manual.* Port Huron, MI: Research Psychologists Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. doi:10.1016/j.jrp.2004.09.009

Le, H., Casillas, A., Robbins, S. B., & Langley, R. (2005). Motivational and skills, social, and self-management predictors of college outcomes: Constructing the student readiness inventory. *Educational and Psychological Measurement, 65*(3), 482–508. doi:10.1177/0013164404272493

Levin, H. M. (2013). The utility and need for incorporating noncognitive skills into large-scale educational assessments. In M. von Davier, E. Gonzelez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 67–86). Dordrecht, Netherlands: Springer.

Liu, M., Bowling, N. A., Huang, J. L., & Kent, T. A. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *TIP: The Industrial and Organizational Psychologist, 51*(1), 32–38.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India, 2*(1), 49–55.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*(1), 61–83. doi:10.1016/j.jrp.2013.09.008

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard-deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences, 84*, 79–83. doi:10.1016/j.paid.2014.08.021

McGrath, R. E., Mitchell, M., & Kim, B. H. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*(3), 450–470. doi:10.1037/a0019216

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. doi:10.1037/a0028085

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72–87. doi:10.1037/1082-989X.8.1.72

Mishkind, A. (2014). *Overview: State definitions of college and career readiness.* Washington, DC: College & Career Readiness & Success Center at American Institutes for Research. Retrieved from http://www.ccrscenter.org/sites/default/files/CCRS%20Defintions%20Brief_REV_1.pdf

Naemi, B., Burrus, J., Kyllonen, P. C., & Roberts, R. D. (2012). *Building a case to develop noncognitive assessment products and services targeting workforce readiness at ETS.* (ETS Research Memorandum RM-12-23). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/workforce_readiness/pdf/rm_12_23.pdf

National Research Council. (2012). *Education for life and work: Developing transferrable knowledge and skill in the 21st century.* Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/download/13398

Ran, S., Liu, M., Marchiondo, L. A., & Huan, J. L. (2015). Difference in response effort across sample types: Perception or reality? *Industrial and Organizational Psychology, 8*(2), 202–208. doi:10.1017/iop.2015.26

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98*(3), 598–616. doi:10.1037/0022-0663.98.3.598

Robbins, S. B., Lauver, K., Le, H., Davis, D., & Langley, R. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261–288. doi:10.1037/0033-2909.130.2.261

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. doi:10.18637/jss.v048.i02

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Steiger, J. H., & Lind, J. M. (1980). *Statistically based tests for the number of common factors.* Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36*(2), 186–212. doi:10.3102/1076998610366263

ACT is an independent, nonprofit organization that provides assessment, research, information, and program management services in the broad areas of education and workforce development. Each year, we serve millions of people in high schools, colleges, professional associations, businesses, and government agencies, nationally and internationally. Though designed to meet a wide array of needs, all ACT programs and services have one guiding purpose—helping people achieve education and workplace success.

ACT.org/research