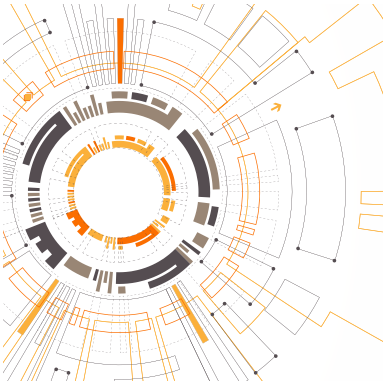


Abstract

Some college students may be disengaged when completing assessments for institutional accountability and improvement. If disengagement is not identified and the resulting data are removed, the validity of score interpretations suffers. Using data gathered from students who completed non-consequential assessments for institutional accountability, we investigated disengagement on a non-cognitive assessment. We demonstrate how we identified students who rapidly responded to items, who "streamlined" answers across items, or who self-reported low effort. We hypothesized that some students would display at least one of these disengagement behaviors and that removing their data would result in scores that better aligned with the assessment's theoretical factor structure. Half the students who self-reported low effort and half the students who streamlined also rapidly responded. The theoretical two-factor structure of the non-cognitive assessment better represented scores after removing disengaged students. We discuss the practicality of selecting a motivation filtering technique to provide more accurate outcome assessment interpretations.

AUTHORS

Katarina E. Schaefer, M.A.
James Madison University

Sara J. Finney, Ph.D.
James Madison University

The Influence of Student Disengagement on a Non-Cognitive Measure: Practical Solutions for Assessment Practitioners

CORRESPONDENCE
Email
schae2ke@jmu.edu

Assessment practitioners generally assume that scores from assessments meaningfully represent some intended construct. That is, the goal is to gather scores with a high degree of validity whether scores are collected under high-stakes conditions (e.g., classroom exams, certification testing, admissions testing) or low-stakes conditions (e.g., institutional accountability assessment, cross-country comparisons) via cognitive assessments where items are scored correct/incorrect (e.g., quantitative reasoning, critical thinking) or non-cognitive assessments with no correct answers (e.g., civic responsibility, global perspectives, growth mindset). However, we should not simply assume that scores represent some intended construct.

As assessment practitioners, we must collect validity evidence to form an argument to support our interpretations of assessment scores, whether those interpretations are shared with accreditors, board of visitors, parents, students, or other stakeholders. *The Standards for Psychological and Educational Testing* (APA, AERA, & NCME, 2014) outline five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequences. Explanations of each source go beyond the scope of this paper.

We will focus on internal structure as a critical piece of evidence impacting score interpretations and how student disengagement on outcomes assessments impacts internal structure.

Disengagement and Low-Stakes Testing

Student disengagement is of concern in low-stakes assessment contexts (e.g., institutional accountability and improvement, cross-country comparisons) and low-stakes data collection (e.g., program evaluations, surveys, data collected for research) contexts (Finn, 2015; Wise, 2006; Wise & DeMars, 2005). In low-stakes contexts, there is no personal consequence to students based on their scores. Hence, some students expend little effort. In turn, scores on both cognitive and non-cognitive assessments may not reflect the construct of interest if a subset of students are disengaged (e.g., Rios et al., 2022; Wise & Kong, 2005).

As assessment practitioners, we must collect validity evidence to form an argument to support our interpretations of assessment scores.

Given that disengagement is an issue in low-stakes assessment contexts, some researchers have attempted to proactively reduce disengagement during testing using techniques such as offering external incentives (e.g., Rios, 2021), increasing test relevance (e.g., Liu et al., 2015), and priming (e.g., Finney et al., 2024). Unfortunately, some methods to reduce disengagement may be costly (external incentives), not possible for certain institutions (increasing test relevance), or may be less effective for certain populations (priming). Furthermore, decreasing disengagement via these proactive methods may not completely eliminate disengagement. Finally, these techniques cannot be applied to data that has already been collected. Thus, in many cases, assessment practitioners can only evaluate if test scores have been contaminated by disengagement, rather than apply a proactive strategy to increase effort. If the disengagement is not addressed, incorrect score interpretations could be made.

Consider higher educational institutions that compute gain scores from low-stakes assessments administered before and after educational programming for accountability or program improvement purposes. If student disengagement is not addressed, the low effort will bias gain scores (Finney et al., 2016; Mathers et al., 2018). These biased value-added estimates can then lead to incorrect inferences about student learning and development. This negative effect of student disengagement is usually a concern for cognitive assessments. However, non-cognitive assessments are popular in higher education to evaluate students' attitudes, perceptions, values, and behaviors (e.g., sense of social belonging, civic engagement, goal orientation, career decision self-efficacy). Surprisingly, there is less discussion surrounding the effect of disengagement on non-cognitive assessments even though these developmental constructs are often the main outcomes of first-year seminars, co-curricular experiences, and student affairs programming.

Thus, the current study focuses on student disengagement when completing non-cognitive assessments. We demonstrate three pragmatic methods that higher education assessment practitioners can use to identify disengagement (streamlining, self-reported effort, and response time) even if the data has already been collected. We then use those methods to gather more robust evidence of the internal structure of scores from the non-cognitive assessment. Finally, we provide practical recommendations for motivation filtering for assessment specialists.

Influence of Disengagement on Internal Structure

The *Standards* state that assessment developers should ensure that items align as intended with the construct(s) of interest and evaluation of this alignment is often accomplished via factor analysis. Consider the Short Grit Scale (Grit-S) used on college campuses; it has some items written to reflect consistency of interest and some items to reflect perseverance of effort (Duckworth & Quinn, 2009). The developer hypothesized that consistency and perseverance influence students' responses to particular items. Structural validity is supported if the intended two-factor structure accounts for the item covariances.

However, beyond the item's content, the factor structure of scores can be influenced by student disengagement, which is an issue for the assessment practitioner who is collecting, analyzing, and interpreting the data. Consider an extreme case of students being completely disengaged on the Grit-S (Duckworth & Quinn, 2009). If students randomly selected responses to the Likert-style items, the two-factor structure would not emerge because the relations among

the item responses would not reflect differences in the two constructs. Thus, the theoretical structure of the responses (which guides the scoring of the measures) would not match the empirical structure of the scores collected by the assessment practitioner (which complicates the reporting and interpretation of scores).

Given the concern of student disengagement on non-cognitive measures, possible influences have been investigated. Barry and Finney (2009) conducted a study where data was collected from college students for accountability and improvement efforts. The assessments were low-stakes for the students (did not affect their GPA), but high-stakes for the institution (scores were reported for accreditation). Barry and Finney varied the assessment contexts to be highly controlled (i.e., proctored, small room) to uncontrolled (i.e., unproctored, remote). Barry and Finney found that for the least controlled context, their data did not align with the measure's theorized internal structure. Although Barry and Finney did not measure disengagement directly, they inferred the internal structure differences across testing contexts were due to differences in disengagement across testing contexts. Fortunately, there are empirical methods to directly identify and address disengagement that are accessible to assessment practitioners.

Methods to Identify Disengagement

Multiple methods exist to identify disengagement. Once identified, people who are disengaged can be removed from the dataset, a technique called "motivation filtering" (e.g., Wise & Kong, 2005). Studies have shown that motivation filtering resulted in higher convergent validity (e.g., Wise & DeMars, 2005; Wise et al., 2004; Wise & Kong, 2005). We discuss and then demonstrate three practical strategies to identify disengagement on non-cognitive assessments and filter non-effortful responses. We present three strategies given the constraints of assessment processes across campuses. Our hope is that at least one of these approaches will be accessible to your assessment context.

Usefulness of Negatively Worded Items to Signal Disengagement

Non-cognitive measures that include both positively and negatively worded items are useful for the identification of extreme disengagement via "streamlined" responses. Streamlined or "longstring" responses occur when people select the same response option for every item, regardless of item wording (e.g., Curran, 2016; Meade & Craig, 2012). For example, five items on a measure may be positively worded (e.g., I am confident in my communication skills), and one item may be negatively worded (e.g., I have poor communication skills). Students rate the items on a scale of 1 to 5, ranging from strongly disagree to strongly agree. It is expected that students who agree or strongly agree with the positive items should disagree or strongly disagree with the negative items. However, disengaged students may not read the items at all and may select the same response (e.g., agree) for the negatively worded items as they did the positively worded items. If some students select the same response option for every item when wording dictates a different style of response, then the factor structure will reflect this via misfit of the intended factor structure. Indeed, assessment specialists employing non-cognitive measures with negatively worded items have used streamlined responses to identify disengagement (e.g., Curran, 2016; Hong et al., 2020; Kupffer et al., 2024; Meade & Craig, 2012).

Negatively worded items are particularly useful to detect extreme disengagement via streamlined responses.

Unfortunately, negatively worded items have a complicated history. Negatively worded items were recommended for inclusion in non-cognitive assessments to help identify acquiescence or disengagement (Bandalos, 2018). However, for decades negatively worded items have been criticized because item valence has been shown to influence the factor structure of scores (Barnette, 2000; Dalal & Carter, 2015; DiStefano & Motl, 2006; Woods, 2006). In particular, factor analyses of non-cognitive assessments with both negatively and positively worded items often result in factors representing item wording (Dalal & Carter, 2015; Ponce et al., 2022). Yet, this resulting factor structure is exactly what one would expect if some students were disengaged; thus, the factor structure signals issues with the quality of responses due to low engagement. However, research also suggests that negatively phrased words can be difficult to comprehend compared to positively phrased words (e.g., Dalal & Carter, 2015; Marsh, 1986, 1996) and that negatively worded items add a level of complexity that may result in misresponse (Dalal & Carter, 2015; Swain et al., 2008). Thus, even when students are engaged, they may struggle to mentally "flip" the meaning of negatively worded items in order to respond in a way that

aligns with the responses to the positively worded items, resulting in factors that represent item valence. Hence, researchers have challenged the use of negatively worded items, with some recommending against their use entirely (e.g., Lindwall et al., 2012; Quilty et al., 2006).

However, we believe that negatively worded items are particularly useful for two reasons. First, they can be used by assessment specialists to detect extreme disengagement via streamlined responses and these invalid responses can be removed. Second, and related to the first reason, negatively worded items can be used to investigate if item-wording factors represent substantively meaningful constructs, ephemeral artifacts of methods effects that are substantively irrelevant, or stable response styles (Marsh et al., 2010).

In short, if disengagement is present on the Rosenberg Self-Esteem Scale and it is not investigated and dealt with, assessment practitioners could draw inaccurate conclusions.

To showcase these two reasons, consider the following example. Assessment practitioners may wish to use the Rosenberg Self-Esteem scale (Rosenberg, 1965) to measure the self-esteem of their students on campus. The 10-item scale has 5 positively worded items and 5 negatively worded items, all intended to measure a single construct. Although the scale is intended to have one factor, numerous studies have challenged the one factor structure in favor of models that account for variance due to item valence (e.g., Lindwall et al., 2012; Marsh et al., 2010; Quilty et al., 2006). Assessment practitioners then need to investigate if the factors that reflect item wording are substantively meaningful, substantively irrelevant artifacts, or response styles. For example, when reviewing studies of the factor structure of the Rosenberg Self-Esteem scale for their meta-analysis, Gnambs et al. (2018) explained that some assessment practitioners interpreted the item-wording factors as positive self-esteem and negative self-esteem. Hence, the item-wording factors may have substantively different meanings. However, this same structure could emerge due to disengagement; thus, the structure would be substantively irrelevant. Disengagement could be manifested in streamlined responses, where some students select the same response option, regardless of the item wording. In this situation, item interrelations would not be fully explained by the one-factor model of self-esteem because responses were also influenced by level of disengagement. An EFA would likely support a two-factor solution based on item valence. If a one-factor CFA model were fitted to the data, correlated residuals between negatively worded items would emerge, necessitating an item-wording method effect factor to reproduce the data adequately. In short, if disengagement is present on the Rosenberg Self-Esteem Scale and it is not investigated and dealt with, assessment practitioners could draw inaccurate conclusions (e.g., meaningful differences between positive and negative self-esteem).

Instead, assessment practitioners could use streamlined responses to identify extreme disengagement. After filtering students who streamline from the dataset, the factor structure could be re-estimated. The factors associated with item valence would dissipate if they mainly reflected this extreme disengagement. In turn, self-esteem would be depicted as a unidimensional construct and the assessment practitioner could report a total self-esteem score for each student. If the item-wording factors remained after removing streamlined responses, then substantively meaningful factors of positive and negative self-esteem or stable response styles may be plausible. This example highlights the usefulness of negatively worded items to detect disengagement, especially for situations where item-wording factors are assumed by some assessment practitioners to be substantively relevant and by others to be substantively irrelevant.

Usefulness of Self-Reported Effort to Signal Disengagement

Not all non-cognitive assessments administered on our campuses contain negatively worded items and streamlined responses are less reliable as a method to detect disengagement without negatively worded items (Curran, 2016). Fortunately, self-report measures of expended effort can be administered by assessment practitioners after an assessment or after a series of assessments. Students who self-report that their motivation is low can be identified and their responses removed from the dataset. For example, higher education assessment practitioners have employed motivation filtering using scores from the effort subscale of the Student Opinion Scale (Thelk et al., 2009). Specifically, a cut-off of 15 has been established, where students who have scores at or below 15 on the effort subscale are identified and removed from the dataset (Swerdzewski et al., 2011).

Self-report measures are not a perfect method to evaluate disengagement. Students may lack engagement when responding to the self-report measure itself (Wise & Kong, 2005). Additionally, students may inaccurately report higher motivation in order to "look better" (Rios et al., 2014), in fear of punishment (Wise, 2020), or students may inaccurately report lower motivation to protect their self-esteem. For example, after a difficult assessment, students may falsely attribute the cause of their poor performance to low levels of effort (Myers & Finney, 2021). Moreover, a recent study found that self-reported effort is not as good of an indicator for disengagement compared to more "behavioral" (e.g., response time, number of clicks) indices (Csányi & Molnár, 2023). Hence, self-reported effort may be most useful for assessment practitioners in conjunction with additional behavioral indices of disengagement.

Usefulness of Response Time to Signal Disengagement

One of the most common behavioral measures of disengagement in higher education and K-12 contexts is response time (e.g., Rios et al., 2014; Wise & Kuhfeld, 2020). Some students answer an item so quickly that they could not have read and processed the item (e.g., Wise & Kong, 2005). Response times can be used as a method to identify and filter responses from disengaged students. For example, assessment specialists have used the mean item response time to identify rapid responses. That is, responses provided in less than an established percent of time (e.g., 20% of the mean time; Wise & Ma, 2012) are considered rapid responses. Responses from students who rapidly responded to more than a certain percentage of items (e.g., rapidly responded on 10% or more of the items) are removed from the dataset.

Unfortunately, there are also issues with using response time to identify disengaged students, particularly in low-stakes higher education testing contexts. Some filtered responses may have been effortful but removed because they were provided quickly. Likewise, some retained responses may have been non-effortful but were not removed because they were associated with long response times (Wise, 2020). Finally, some data collections do not allow for timing data to be collected at the item-level or the measure-level (e.g., some commercially available measures used in higher education for accountability do not provide the institution with timing data).

Streamlining, self-reported effort, and response time are effective means for assessment practitioners to identify student disengagement. However, the three strategies measure different manifestations of disengagement; thus, they do not identify the exact same sample of disengaged students. For example, although we know that disengagement is related to test performance (Rios et al., 2022), self-reported effort has a smaller correlation with performance (r = .33) compared to response time (r =.72), suggesting that the two reflect different manifestations of disengagement (Silm et al., 2020). Moreover, the correlation between self-reported effort and response time effort (e.g., Wise & Kong, 2005) varies between small to moderately large (r = .28, Akhtar & Firdiyanti, 2023; r = .13, Csányi & Molnár, 2023; r = .61, Rios et al., 2014; r = .25, Wise & Kong, 2005). Swerdzewski et al., (2011) found that response time effort and self-reported effort agreed (i.e., flagged the same individuals as having low motivation) for 66.01% of the disengaged students. Additionally, one study found that streamlined responses correlate in a small but positive way with self-reported effort (.16) but had a very small (-.06) correlation with response time in minutes (Kupffer et al., 2024). Thus, although it is expected that some students who self-report having low effort will also streamline or rapidly guess, there will be students who do not exhibit multiple types of disengagement. Indeed, using a "hurdle approach," where multiple methods to detect disengagement are used simultaneously, is recommended to identify different types of disengagement (e.g., Curran, 2016; Goldammer et al., 2020; Meade & Craig, 2012).

Self-reported effort may be most useful for assessment practitioners in conjunction with additional behavioral indices of disengagement.

The Current Study

Using data collected for institutional accountability and improvement purposes, we investigated the effect of different disengagement types on the internal structure of a non-cognitive assessment. Three disengagement identification methods were used. Students self-reported if they had low effort while completing assessments (e.g., Swerdzewski et al., 2011). "Rapid responders" were students who responded so quickly they could not have read the item (e.g., Wise & DeMars, 2005; Wise & Kuhfeld, 2020). "Streamliners" were students who

consistently selected the same response option (e.g., Curran, 2016; Hong et al., 2020; Steedle et al., 2019). The impact of disengagement on the factor structure of scores was estimated and compared across disengagement methods with the goal that at least one of these approaches will be useful and accessible for assessment practitioners. Specifically, this study addressed two hypotheses:

1) There would be at least a moderate proportion of disengaged students identified as being disengaged by at least one method (streamlining, self-report, and response time). We did not expect that all disengaged students would be identified by all three methods, given that the three indicators of disengagement reflect different manifestations of disengagement. That is, we expected that a moderate number of students who self-reported having low-effort would also rapidly respond, given the moderate relationship between the two indicators. We expected the fewest disengaged students to be flagged using the streamline method. Moreover, of those who did streamline, only a small proportion would also self-report having low motivation and would rapidly respond, given that streamlining inconsistently aligns with other indicators of disengagement (e.g., Goldammer et al., 2020; Hong et al., 2020).

The impact of disengagement on the factor structure of scores was estimated and compared across disengagement methods with the goal that at least one of these approaches will be useful and accessible for assessment practitioners.

2) After removing disengaged students, the internal structure of the scores from a non-cognitive measure should be less contaminated by disengagement; thus, a CFA model reflecting the intended internal structure of the scores would fit the data better. This improved fit would be evidenced via global fit indices and reduced correlation residuals between the negatively worded items. If all disengagement methods improve model-data fit, assessment practitioners can use motivation filtering with any of these methods to improve the validity of their score interpretations.

Method

Participants and Procedure

Our mid-size (approximately 20,000 students) southeastern US university uses low-stakes assessments to evaluate outcomes of our general education programming (e.g., quantitative reasoning) and university-wide initiatives (e.g., civic engagement). Assessments are administered to incoming students in the fall and advanced students in the spring. Although every student was required to complete a series of assessments taking approximately two hours, scores have no personal impact on students (e.g., no impact on grades, awards, opportunities).

Data from incoming first-year students were collected in the fall of 2021 under low-stakes conditions. All students completed a series of cognitive and non-cognitive assessments. Only students who completed the non-cognitive assessment of interest, who consented to having their data used for research purposes, and who were over the age of 18 were included in the analysis, which resulted in 3,169 students. Assessments were administered online and unproctored via Qualtrics. Students had a multiweek window during which they were required to complete the assessments.

Measure

The Attitudes Towards Communication (ATC) assessment is the primary assessment of interest in the current study. The 11-item non-cognitive assessment has two subscales: willingness to communicate (6 items) and confidence in communication (5 items). These subscales are intended to measure "key communication concepts for undergraduate college students" (Williams et al., 2014). Items on the ATC were developed to assess affective components of communication and written to align with the National Communication Association standards.

We believed that some students would put forth little effort when completing the ATC assessment for two reasons. First, students tend to put forth less effort as assessments get longer (e.g., Pastor et al., 2019) or are later in the testing session (e.g., Finney & McFadden, 2023). Although the 11-item ATC assessment could be considered relatively short and less cognitively demanding than say a math assessment, we were concerned about low-effort on this assessment because it was the second assessment in a series of assessments. Thus, some students may be fatigued later in the testing session, which would result in them putting forth little effort on the ATC assessment. Second, low-stakes testing contexts are associated with lower effort than

high-stakes testing contexts due to the lower perceived importance of these tests to the students (e.g., Finney et al., 2018; Satkus & Finney, 2021). Even if the ATC assessment was placed first in the series of assessments, we would still advise practitioners to investigate effort on the ATC due to the non-consequential nature of the testing context, regardless of the assessment's length or content. Finally, no methods to proactively increase test-taking motivation (e.g., incentives, increasing test relevance, priming) were used in the current study.

ATC Assessment Theorized Internal Structure

The ATC assessment was designed with two intentionally distinct subscales: willingness and confidence. Willingness items measure students' openness to communication. Confidence items measure communication self-efficacy. A two-factor structure with no cross-loadings was expected. All ATC items were responded to using a 5-point Likert scale: 1 (*Strongly Disagree*), 2 (*Disagree*), 3 (*Undecided*), 4 (*Agree*), and 5 (*Strongly Agree)*. Note, one of the negatively worded items was theorized to reflect willingness (item 3) whereas the other negatively worded item was theorized to reflect confidence (item 8). Higher scores represent higher willingness or confidence to participate in speech performance. Cronbach's alpha for the willingness scores ($\alpha = 0.78$) and confidence scores ($\alpha = 0.74$) was adequate.

Indicators of Disengagement

Self-Report Measure

After approximately two hours of completing assessments, students completed a self-report measure of effort. More specifically, students completed a cognitive test, followed by the ATC scale, and then responded to the Student Opinion Scale (SOS) (Pastor et al., 2023; Thelk et al., 2009). The SOS contains a five-item subscale intended to measure expended effort for the total testing session. Thus, effort on the SOS reflects not only effort on the non-cognitive ATC scale, but also effort on the cognitive test as well. Effort scores from the SOS range from 5 (no effort) to 25 (highest effort). Filtering scores from disengaged students was accomplished using a cutoff score of 15 on the effort subscale, as was done in previous studies using this self-report measure (e.g., Swerdzewski et al., 2011). ATC scores from students who scored at or below 15 on the effort subscale were removed from the filtered datasets.

Response Time

Response times were not available for the non-cognitive ATC measure but were available for the cognitive test taken just prior to the ATC. This use of response time on a previous task was justified because rapid response behavior tends to increase throughout a series of tests (e.g., Pastor et al., 2019) and measures of motivation on one task have been used to make inferences about motivation on an accompanying task (Zamarro et al., 2019). The current study used the normative threshold setting method (NT20) to identify students who rapidly guessed on the previous cognitive test (Wise & Ma, 2012). First, the mean response time for each cognitive item was calculated. Then, if the response time for the item was lower than 20% of the mean response time for that item, the item response was flagged as a rapid response. Finally, if a student rapidly responded on more than 10% of the cognitive items, they were identified as a rapid responder (e.g., Rios et al., 2017; Wise & DeMars, 2010). ATC scores from students who rapidly responded on the cognitive test prior to the ATC were removed from the filtered datasets.

The ATC assessment was designed with two intentionally distinct subscales: willingness and confidence.

Streamlining Responses

Students were categorized as streamliners if they selected the same response option (e.g., all "strongly agree," all "disagree") to all items on the non-cognitive measure prior to reverse scoring the two negatively worded items. Students were not categorized as streamliners if they selected "undecided" for all response options, given that selecting "undecided" is a valid option for both positively and negatively worded items. Of those students who responded to the ATC assessment, only 2.2% (70 out of 3,169) selected "undecided" for all items on the confidence subscale, 3.1% (98 out of 3,169) selected "undecided" for all items on the willingness subscale, and 2.1% (66 out of 3,169) selected "undecided" for all items on the ATC (across both subscales).

These students were not categorized as streamliners. ATC scores from students who exhibited streamlining on this non-cognitive measure were removed from the filtered datasets.

Results

Number of Students Identified as Disengaged Across the Three Methods

Frequencies and proportion of students identified as disengaged by each of three methods are displayed in Table 1. Unfortunately, approximately 24% of all students (745 students out of the total sample of 3,169) were disengaged in some way (streamlined, rapidly responded, or self-reported low effort). Yet, we felt fortunate that we were able to actually identify this disengagement instead of assuming it did not exist. Of those who displayed disengagement, most rapidly responded on the previous cognitive test or self-reported low effort. Fewer students streamlined, although streamliners still accounted for a meaningful, although relatively small, number of disengaged students (141 students).

Table 1
Proportion of Total Students who were Disengaged

Disengagement Type	N	% (out of 3,169)
Streamliners	141	4.4%
Rapid Responders	488	15.4%
Low Self-Reported Effort	425	13.4%
Total Disengaged	745	23.5%

Note. Total disengaged does not equal the sum of all disengagement types in the table. Some students used more than one disengagement type.

To address our first hypothesis, we computed the proportion of students who streamlined, rapidly responded, or self-reported having low effort (Table 2). Nearly half (≈ 41%) of those who streamlined (n = 141) also rapidly responded (n = 58). Nearly half (≈ 42%) of those who self-reported having low effort (n = 425) also rapidly responded (n = 179). Very few (≈ 6%) who streamlined (n = 141) also self-reported having low effort (n = 8). Finally, very few (≈ 4%) of those who exhibited at least one of the disengagement types (n = 745) used all three disengagement types (n = 32). Figure 1 displays a proportional Venn diagram of students who exhibited different disengagement types.

Confirmatory Factor Analysis: Model Fit

We addressed our second hypothesis using confirmatory factor analysis (CFA). CFA was used to assess model-data fit for five different data sets: 1) unfiltered dataset containing all students, 2) filtered dataset without students who streamlined, 3) filtered dataset without students who rapidly responded, 4) filtered dataset without students who self-reported having low effort, and 5) filtered dataset without students who displayed any disengagement type. Due to there being only small, nuanced differences in the CFA models between the three samples in which only one type of disengagement was removed (streamlining only, rapidly responding only, and self-reporting only), we did not conduct analyses on samples in which two types of disengagement were removed (streamlining and rapidly responding, streamlining and self-reporting, and rapidly responding and self-reporting).

Approximately 24% of all students were disengaged in some way.

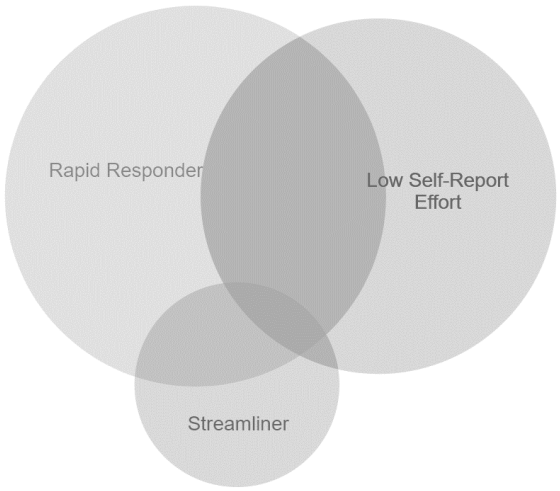
CFA analyses were conducted using Mplus version 8.6. Items were approximately normally distributed other than Item 2 having kurtosis of 5.56. We compared the results of models estimated using maximum likelihood (ML) estimation and maximum likelihood with the Satorra-Bentler adjustment (MLMV) (Finney et al., 2016). Due to the negligible difference in results and inferences when comparing ML and MLMV, we determined that the data were sufficiently normally distributed, and ML results are reported.

Statistical analyses were not used to compare the results across the different filtered samples. Instead, results across the different samples were compared via both global and local fit indices. The CFI ranges from 0 to 1 with higher values indicating better model-data fit. The

Table 2
Proportion of Disengaged Students by Disengagement Method

Disengagement Method	N	% (out of 745)
Streamliner ONLY	43	5.8%
Rapid Responder ONLY	219	29.4%
Low Self-Reported Effort ONLY	206	27.7%
Streamliner & Rapid Responder	58	7.8%
Streamliner & Low Self-Reported Effort	8	1.1%
Rapid Responder & Low Self-Reported Effort	179	24.0%
Streamliner, Rapid Responder, & Low Self-Reported Effort	32	4.3%
Total	745	100.0%

Figure 1
Overlap Between Disengagement Types



Note. Proportions are shown approximately to scale. Nearly half (≈ 41%) of those who streamlined also rapidly responded. Nearly half (≈ 42%) of those who self-reported low effort also rapidly responded. Few (≈ 6%) who streamlined also self-reported having low effort.

RMSEA and SRMR range from 0 to 1 where lower values indicate better model-data fit. As shown in Table 3, Models 2 to 5 (filtered data sets) fit the data better in a global sense than Model 1 (unfiltered data set). Sample 4 (only low self-reported effort removed) resulted in fit index values that were close in magnitude to the fit indices of Sample 1 (unfiltered). Sample 4 also resulted in the lowest CFI relative to all other samples. The largest differences in fit index values were between Sample 1 (unfiltered) and Sample 5 (all disengagements removed).

Correlation residuals, which represent the discrepancy between the corresponding observed and model-implied item-level correlations, were used to assess local model-data misfit (Bandalos & Finney, 2019). If a model fits perfectly, the correlation residuals are zero. Comparisons were made between the size of the residual correlation between the two negatively worded items, as well as the number of residual correlations over |0.10| and |0.15|. As expected, there were fewer correlation residuals over |0.10| and |0.15| in the filtered data sets compared to the unfiltered data set (see Table 4).

Table 3
Comparison of Global Fit Indices for Two-Factor Model Fit to Unfiltered and Filtered Samples

Sample Type	*n*	$\chi 2$	*df*	CFI	RMSEA	SRMR
Sample 1: Unfiltered	3169	1190.16*	43	0.886	0.092	0.058
Sample 2: Streamliners Removed	3028	949.42*	43	0.904	0.083	0.052
Sample 3: Rapid Responders Removed	2681	745.26*	43	0.895	0.078	0.049
Sample 4: Low Self-Reported Effort Removed	2744	926.52*	43	0.878	0.087	0.054
Sample 5: Any Disengagements Removed	2424	653.39*	43	0.898	0.077	0.047

The correlation residual between the two negatively worded items decreased from the unfiltered data set to the filtered data sets. That is, for the unfiltered data, even after accounting for the constructs of interest (willingness and confidence), there was still a non-negligible (residual) correlation between the negatively worded items, caused by students' disengagement. However, this construct-irrelevant relation diminished when disengagement was filtered from the dataset, as hypothesized. There are two notable changes in the correlation residuals. First, the correlation residual from the unfiltered data set dropped from 0.21 to 0.03 in the data set when all disengagement types were filtered. This difference is substantial. Second, the correlation residual dropped from 0.21 to 0.10 when only streamliners (which were only 4.4% of the sample) were removed. These results provide evidence of the impact of disengagement on the factor structure of scores and, in turn, structural validity evidence.

Table 4
Comparison of Local Fit (Correlation Residuals) for Unfiltered and Filtered Samples

Sample	Correlation Residual Between the Two Negatively Worded Items	Frequency of Correlation Residuals	
		> |.10|	> |.15|
Sample 1: Unfiltered	0.208	9	3
Sample 2: No Streamliners	0.101	6	1
Sample 3: No Rapid Responders	-0.054	6	3
Sample 4: No Low Self-Reported Effort	0.185	7	2
Sample 5: No Disengagements	0.034	4	1

Discussion

The current study demonstrated three different practical strategies that identified students who were not engaged when completing low-stakes assessments. Moreover, we used these three methods to remove invalid responses from the dataset, which in turn positively influenced the structural validity of the responses. Results of our research questions and practical implications for assessment practitioners are discussed.

The correlation residual from the unfiltered data set dropped from 0.21 to 0.03 when all disengagement types were filtered.

Disengaged Students Identified via Different Methods

We hypothesized there would be some students who were identified by at least one disengagement method and that some students would be identified by more than one disengagement method. Our results supported our first hypothesis. We found that about half of those who self-reported having low effort also rapidly responded. This finding aligns with results of studies that found moderate correlations between self-reported effort and response

time (Akhtar & Firdiyanti, 2023; Csányi & Molnár, 2023; Rios et al., 2014; Wise & Kong, 2005) and aligns with results of a study that directly estimated the percent of students (66.01%) who self-reported low effort and who rapidly guessed (Swerdzewski et al., 2011). Interestingly, about half of those who streamlined also rapidly responded. This finding aligns with research that suggests both of these behavioral indicators of disengagement (rapid guessing, Akhtar & Firdiyanti, 2023; streamlining, Hong et al., 2020) are good indicators of disengagement.

Finally, few students who streamlined also self-reported having low effort. We believe that the self-reported effort measure could have been affected by streamlining. If students streamlined through the self-report effort measure, they might not be captured as having self-reported low effort. Moreover, self-reported effort is for the whole testing session, not the ATC specifically. We may expect more alignment between students who self-report low effort for the ATC and who streamline responses to the ATC items.

The Effect of Disengagement on Factor Structure

We hypothesized that removing responses from disengaged students would result in improved model-data fit (i.e., theorized internal structure would better match the empirical internal structure and thus the recommended scoring of the measure could be employed). Our results supported our hypothesis. When responses from disengaged students were removed from the dataset, model-data fit improved, regardless of the motivation filtering method chosen. In fact, the correlation residual between the negatively worded items was close to zero after removing students exhibiting any type of disengagement. In the current study, model-data fit indices did not meet recommended cut-off values. However, having good model-data fit was not the purpose of the current study. The purpose of the current study was to take steps to improve the validity of score interpretations by removing the influence of disengagement from the internal structure of scores via methods that are practical and accessible to assessment practitioners. Filtering by any disengagement method improved model-data fit, with rapid guessing and streamlining producing better results. If possible, assessment practitioners should use multiple techniques to identify disengaged students. However, practitioners can rest assured that using at least one method will still result in improved interpretations via improved internal structure.

Filtering by any disengagement method improved model-data fit, with rapid guessing and streamlining producing better results.

Some researchers note that streamlined responses are not always a good indicator of disengagement (Goldammer et al., 2020). Moreover, the detection of streamlining requires the inclusion of at least one negatively worded item, even though inclusion of negatively worded items has been admonished by some (e.g., Dalal & Carter, 2015) and may not be possible to include in all assessment contexts. However, in our study, the effect of item valence on the factor structure of scores was substantially reduced after removing a small number of students who streamlined (only 4.4% of the sample). In our sample, streamlining represented an extreme, flagrant form of disengagement used by a very small number of students. As a consequence, we believe that streamlining (which requires negatively worded items) can be used as an effective means to identify disengagement and improve the factor structure of scores without removing a large number of students and thus retain better generalizability of the scores.

Limitations

There are some limitations to the current study. First, response time was gathered using an adjacent assessment. Ideally, response time should be gathered on the assessment of interest, rather than on an adjacent assessment. With that said, we realize other assessment practitioners may encounter this same issue given response time is not always available for all measures (some commercial measures used in higher education do not report response time). Thus, our work can be used as a reference for those who find themselves unable to collect response time on their measure of interest.

Second, the current study did not vary the content nor the length of the non-cognitive measure. Thus, the generalizability of our results may only extend to measures of similar length and content. Future studies may vary the length and/or content of the non-cognitive measure to investigate the impact of low motivation on the factor structure of scores.

Third, assessment specialists caution the use of motivation filtering when disengagement is related to ability on cognitive tests (Rios et al., 2017). When disengagement is related to ability, filtering may result in a less generalizable sample because low-ability students are removed from the sample at a disproportionate rate. In short, regardless of whether a test is cognitive or non-cognitive, when motivation filtering is used, the sample characteristics may change when compared to unfiltered data (e.g., change in proportion of high-performing students, change in demographics). Thus, filtering scores from low-motivated students for any type of measure may alter the population of students to whom the test scores may generalize. Thus, higher education assessment practitioners should also focus on increasing engagement a priori. Proactive strategies such as offering external incentives (e.g., Rios, 2021), increasing test relevance (e.g., Liu et al., 2015), and priming (e.g., Finney & McFadden, 2023; Finney & Pastor, 2025) have increased test-taking motivation and reduced the percentage of responses needing to be filtered. If possible, we encourage coupling these proactive strategies to mitigate disengagement with the strategies we showcased in this study. We hope the current study provides assessment practitioners in higher education low-stakes testing with accessible tools to address disengagement, with an understanding of the limitations of the generalizability of our results.

Coupling proactive strategies to mitigate disengagement with the strategies we showcased can increase engagement.

Implications for Higher Education Assessment Practitioners

In closing, filtering invalid responses from disengaged students by any method improved the factor structure of scores and thus the validity of score interpretations. Given the popularity of non-cognitive assessments in higher education, we recommend that assessment practitioners 1) select the disengagement identification technique(s) that is most accessible to them and 2) as a matter of routine, investigate the amount of disengagement present during their collection of outcomes assessment data. Understanding and tackling the issue of student disengagement on outcomes assessments allows for more accurate interpretations of student learning and development data that can be used for accountability reporting and programmatic improvement.

References

Akhtar, H., & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*, *106*. https://doi.org/10.1016/j.lindif.2023.102323

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Bandalos, D. L., & Finney, S. J. (2019). Factor analysis: Exploratory and confirmatory. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd ed., pp. 98–122). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315755649-8

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361-370. https://doi.org/10.1177/00131640021970592

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment*, *3*, 1–15. https://eric.ed.gov/?id=EJ1062735

Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*. https://doi.org/10.1016/j.lindif.2023.102340

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

Dalal, D. K., & Carter, N. T. (2015). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 112–132). Routledge/Taylor & Francis Group.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1–17. https://doi.org/10.1002/ets2.12067

Finney, S. J., DiStefano, C., & Kopp, J. P. (2016). Overview of estimation methods and preconditions for their application with structural equation modeling. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (1st ed., pp. 135 - 165). Hogrefe. https://psycnet.apa.org/record/2016-44627-008

Finney, S. J., & McFadden, M. E. (2023). Examining the question-behavior effect in low-stakes testing contexts: A cheap strategy to increase examinee effort. *Educational Assessment*, *28*(4), 211-228. https://doi.org/10.1080/10627197.2023.2222588

Finney, S. J., Myers, A. J., & Mathers, C. E. (2018). Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing*, *18*, 297 – 322. https://doi.org/10.1080/15305058.2017.1396466

Finney, S. J., Schaefer, K. E., & McFadden, M. E. (2024). Priming examinees to give good effort: Differential utility across gender identity. *The Journal of Experimental Education.* https://doi.org/10.1080/00220973.2024.2310678

Finney, S. J. & Pastor, D. A. (2025). Priming non-compliant students to expend test-taking effort: How many primes are needed? *Journal of Experimental Education*. https://doi.org/10.1080/00220973.2025.2459392

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the Rosenberg self-esteem scale: A cross-cultural meta-analysis. *Zeitschrift für Psychologie*, *226*(1), 14-29. https://doi.org/10.1027/2151-2604/a000317

Goldammer, P., Annen, H., Stockli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), https://doi.org/10.1016/j.leaqua.2020.101384

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2). https://doi.org/10.1177/0013164419865316

Kupffer, R., Frick, S., & Wetzel, E. (2024). Detecting careless responding in multidimensional forced-choice questionnaires. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644231222420

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*(2), 196-204. https://doi.org/10.1080/00223891.2011.645936

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49. https://doi.org/10.1037/0012-1649.22.1.37

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg self-esteem scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*(2), 366–381. https://doi.org/10.1037/a0019225

Mathers, C., Finney, S. J., & Hathcoat, J. D. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education*, *43*(8), 1211-1227. https://doi.org/10.1080/02602938.2018.1443202

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

Myers, A. J., & Finney, S. J. (2021). Does it matter if examinee motivation is measured before or after a low-stakes test? A moderated mediation analysis. *Educational Assessment*, *26*(1), 1-19. https://doi.org/10.1080/10627197.2019.1645591

Pastor, D., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189-212. https://doi.org/10.1080/10627197.2019.1615373

Pastor, D., Patterson, C., & Finney, S. J. (2023). Development and internal validity of the Student Opinion Scale: A measure of test-taking motivation. *Journal of Psychoeducational Assessment*, *41*(2), 209-225. https://doi.org/10.1177/07342829221140957

Ponce, F. P., Irribarra, D. T., Vergés, A., & Arias, V. B. (2022). Wording effects in assessment: Missing the trees for the forest. *Multivariate Behavioral Research*, *57*(5), 718-734. https://doi.org/10.1080/00273171.2021.1925075

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg self-esteem scale method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(1), 99-117. https://doi.org/10.1207/s15328007sem1301_5

Rios, J. A. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356-373. https://doi.org/10.1080/10627197.2022.2110465

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74-104.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *161*, 69-82. https://doi.org/10.1002/ir.20068

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press. https://www.jstor.org/stable/j.ctt183pjjh

Satkus, P. & Finney, S. J. (2021). Antecedents of examinee motivation during low-stakes tests: Examining the variability in effects across different research designs. *Assessment and Evaluation in Higher Education*, *46*, 1065-1079. https://doi.org/10.1080/02602938.2020.1846680

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review, *Educational Research Review*, *31*. https://doi.org/10.1016/j.edurev.2020.100335

Steedle, J., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social-emotional learning competencies. *Educational Measurement Issues and Practice*, *38*, 101-111.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129–151. https://www.jstor.org/stable/27798135

Williams, L. M., & Horst, S. J., & Sundre. D. L. (2014). Test of oral communication skills, version 2: TOCS-II test manual. Harrisonburg, VA: Center for Assessment and Research Studies and Madison Assessment. https://www.madisonassessment.com/assessment-testing/test-of-oral-communication-skills

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, San Diego, CA, United States.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163 –183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. Margolis & R. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pp. 150 – 64). Routledge. https://doi.org/10.4324/9781351064781-11

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a cat item pool: The normative threshold method* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, Vancouver, Canada. https://www.nwea.org/resources/setting-response-time-thresholds-cat-item-pool-normative-threshold-method/

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 189-194. https://doi.org/10.1007/s10862-005-9004-7

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, *13*(4), 519-552. https://doi.org/10.1086/705799