Wright State University

CORE Scholar

Browse all Theses and Dissertations | Theses and Dissertations

2022

Investigating the Efficacy of Novel Measures of Careless Responding to Tests

Mark Christopher Ramsey
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

 Part of the Industrial and Organizational Psychology Commons

Repository Citation

Ramsey, Mark Christopher, "Investigating the Efficacy of Novel Measures of Careless Responding to Tests" (2022). *Browse all Theses and Dissertations*. 2631.
https://corescholar.libraries.wright.edu/etd_all/2631

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

INVESTIGATING THE EFFIACY OF NOVEL MEASURES OF CARELESS RESPONDING TO TESTS

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

by

MARK CHRISTOPHER RAMSEY
B.A., The Ohio State University, 2020

2022
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

June 10th, 2022

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Mark Christopher Ramsey ENTITLED Investigating the Efficacy of Novel Measures of Careless Responding to Tests BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

_____
Nathan Bowling, Ph.D.

Thesis Director

_____
Scott Watamaniuk Ph.D.
Graduate Program Director

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

Committee on Final Examination:

_____
Nathan Bowling, Ph.D.

_____
Corey Miller, Ph.D.

_____
Anthony Gibson, Ph.D.

_____
Barry Milligan, Ph.D.
Dean of the Graduate School

ABSTRACT

Ramsey, Mark Christopher. Department of Psychology, Wright State University, 2022. Investigating the efficacy of Novel Measures of Careless Responding to Tests

Research has demonstrated that careless responding (CR) threatens the construct validity of measures (see Huang et al., 2015; Wise & Kong, 2005). Researchers have developed and studied many measurement approaches to capture CR in surveys, with different survey measures compensating for the practical or empirical limitations of other measures. This research is distinguished from ability test CR research because ability tests are fundamentally different from surveys. Within ability tests, CR research has focused only on response time and self-report measures of CR, both of which carry limitations. The former is inflexible because the index necessitates item-level response time information, and therefore cannot be used in pen-and-paper tests or online tests without such item-level information available. The latter index is plagued by theoretical and empirical shortcomings. Thus, the purpose of my study is to find a comparably valid and more flexible approach through testing the efficacy of five survey CR measurement approaches, namely the infrequency approach, instructed-response approach, the consistency approach, the self-report approach, and long-string analysis, in capturing CR in tests. In a sample of 291 undergraduate students, I found strong support for using the infrequency approach to assess careless responding, weak support for the instructed-response approach and long-string analysis, and no support for the self-report or consistency approaches.

TABLE OF CONTENTS

Page

I. INTRODUCTION....1

Tests and Surveys....2

Careless Responding....2

Prevalence and Effects of Careless Responding....3

Prevention of Careless Responding....8

Measurement of Careless Responding....11

Careless Responding Measurement Approaches Used in Tests....13

Nomological Network....19

Careless Responding Measurement Approaches Used in Surveys....23

II. METHOD....48

Design....48

Manipulations....49

Measures....50

III. RESULTS

Preliminary Analyses....59

Hypotheses 1a-1f....67

Hypotheses 2a-2f....69

Hypotheses 3a-3f....71

Hypotheses 4a-4f.....73

Hypotheses 5a-5f.....75

Hypothesis 6.....77

IV. DISCUSSION.....80

Performance of Novel Measures under Nomological Network.....84

Practical Implications.....91

Limitations.....92

Future Research.....93

Practical Recommendations..... 94

V. References.....96

Appendices.....163

LIST OF FIGURES

Figure Page

1. Plot of Significant Interaction Effect of MAT RTE on the Relationship between Self-Reported ACT scores and MAT Performance.....121
2. Plot of Significant Interaction Effect of SILS-V RTE on the Relationship between Self-Reported ACT scores and MAT Performance.....122
3. Plot of Significant Interaction Effect of MAT RTE on the Relationship between Self-Reported GPA scores and MAT Performance.....123
4. Plot of Significant Interaction Effect of MAT Infrequency on the Relationship between Self-Reported college ACT scores and MAT Performance.....124
5. Plot of Significant Interaction Effect of Long-String Total on the Relationship between Self-Reported college ACT scores and MAT Performance.....125
6. Plot of Significant Interaction Effect of Long-String Total on the Relationship between Self-Reported college ACT scores and SILS-V Performance.....126

LIST OF TABLES

Table Page

1a. Table Describing Careless Responding Indices.....111

1b. Table Describing Careless Responding Indices.....112

1c. Table Describing Careless Responding Indices.....113

2. Nomological Network Table for Capturing Careless Responding to Tests.....114

3. Order of Assessments.....115

4a. Means, Standard Deviations, and Correlations with Response Time Effort.116

4b. Correlations between Careless Responding Indices.....117

4c. Correlations with Low-Stakes Test Performance.....118

5. Performance of RTE and SOS within Nomological Network.....120

6a1. Regression Results using MAT Performance as the Criterion.....127

6a2. Regression Results using SILS-V Performance as the Criterion.....128

6b1. Regression Results using MAT Performance as the Criterion.....129

6b2. Regression Results using SILS-V Performance as the Criterion.....130

6c1. Regression Results using MAT Performance as the Criterion.....131

6c2. Regression Results using SILS-V Performance as the Criterion.....132

6d1. Regression Results using MAT Performance as the Criterion.....133

6d2. Regression Results using SILS-V Performance as the Criterion.....134

6e1. Regression Results using MAT Performance as the Criterion.....135

6e2. Regression Results using SILS-V Performance as the Criterion.....136

7a1. Regression Results using MAT Performance as the Criterion.....137

7a2. Regression Results using SILS-V Performance as the Criterion.....138

7b1. Regression Results using MAT Performance as the Criterion.....139

7b2. Regression Results using SILS-V Performance as the Criterion.....140

7c1. Regression Results using MAT Performance as the Criterion.....141

7c2. Regression Results using SILS-V Performance as the Criterion.....142

7d1. Regression Results using MAT Performance as the Criterion.....143

7d2. Regression Results using SILS-V Performance as the Criterion.....144

7e1. Regression Results using MAT Performance as the Criterion.....145

7e2. Regression Results using SILS-V Performance as the Criterion.....146

8a. ANOVA Results Using MAT Infrequency as Criterion.....147

8a2. ANOVA Results Using SILS-V Infrequency as Criterion.....148

8b. ANOVA Results Using MAT Instructed-Response as Criterion.....149

8b2. ANOVA Results Using SILS-V Instructed-Response as Criterion.....150

8c. ANOVA Results Using Psychometric Synonyms as Criterion.....151

8d. ANOVA Results Using Long-String Total as Criterion.....152

8e. ANOVA Results Using Diligence as the Criterion.....153

9a. ANOVA Results Using SILS-V Infrequency as Criterion.....154

9b. ANOVA Results Using SILS-V Instructed-Response as Criterion.....155

9c. ANOVA Results Using Psychometric Synonyms as Criterion.....156

9d. ANOVA Results Using Long-String Total as Criterion.....157

9e. ANOVA Results Using Diligence as the Criterion.....158

10. Summary of Results Concerning Nomological Network.....159

11. Positive Participant Reactions to Open-Response Item.....160

12. Proportion of Correct Responses to Infrequency Items.....161

13. Proportion of Correct Responses to Instructed-Response Items.....162

Tests and Surveys

Survey and test items require effort for completion. In a typical survey item, Tourangeau (1984) described that participants have to go through four stages of effortful cognitive processing (Tourangeau, 1984). Specifically, participants must attend closely to and understand the item content, recall relevant information, use that information to make a judgment, and implement that judgment into a response (Tourangeau, 1984). Researchers and practitioners have often presupposed that respondents will provide proper effort and attention in responding to items. Challenging this presupposition, recent research has found that often participants forgo this effortful processing and instead respond carelessly to survey assessments (Huang et al., 2012; Meade & Craig, 2012; Maniaci & Rogge, 2014).

Test items are fundamentally different from survey items. Test items assess performance in a given task, whereas survey items assess participants through inquiring about beliefs, attitudes, and past experience. Additionally, tests typically contain items that have incorrect and correct answers that are scored dichotomously. Conversely, survey items do not typically have correct and incorrect responses.

Test items also require effort for completion. To elaborate, Sternberg (1985) outlines 7 effortful processes of inductive reasoning that participants sequentially engage in to solve analogy problems. Respondents first encode the item material, identify a relation between concepts, and then identify a shared-rule between these concepts.

Following this identification, respondents then generalize that shared-rule to another set of concepts, compare and discriminate between response options based on which response fits the rule, and then provide a response to the item.

Often, researchers have presupposed that participants are attentive and effortful in their responding when administering tests and survey items. However, research has suggested clearly that participants often engage in careless responding (CR; see Meade & Craig, 2012 for CR in surveys; see Wise & Kong, 2005 for CR in low-stakes tests). Within the first section of my work, following a discussion of the definitions of CR, I discuss research concerning the effects and prevalence of CR. Following that discussion, I analyze several different measurement approaches of CR. Within the analysis, I argue that these novel measurement approaches can be adapted to successfully measure CR to tests. Adaptation is necessary given that test items fundamentally differ from survey items. I then describe a nomological network that I use to investigate the validity of these novel measures.

Careless Responding

Careless responding (CR) is defined in the literature in one of two ways. Huang et al. (2012) defined careless responding as a set of responses in which the "respondent answers" with "low or little motivation" to comply with "instructions, correctly interpret item content and provide accurate answers". This definition implies that low motivation, as opposed to fatigue for example, primarily causes CR. Conversely, Meade and Craig

(2012) defined CR as the failure to respond with regard to item content (Nichols et al., 1989). The Meade and Craig (2012) definition is simpler and does not imply that any specific phenomenon underlies CR. This latter point is especially important because there is not enough evidence to assert confidently that low motivation, or any single phenomenon, is the primary cause behind CR. Therefore, when I use the term CR, I am using the Meade and Craig (2012) definition.

CR has different labels in published research. CR is also known as insufficient effort responding (see Huang et al., 2012), participant inattention (see Maniaci & Rogge, 2014), content nonresponsetivity (see Nichols et al., 1989), and random responding (see Beach 1989 and Berry et al., 1992). Of the five labels, insufficient effort responding and random responding assume that low motivation is the chief cause behind CR and that careless response patterns are random in nature respectively. Currently, there is an absence of evidence to support the former assertion. Concerning the latter assertion, research has revealed that not all careless response patterns are random in nature (see DeSimone et al., 2018). Of the remaining three, the term careless responding is simpler in nature than participant inattention and content nonresponsetivity. Hence, I have used CR in this work.

Prevalence and Effects of Careless Responding

Prevalence of Careless Responding. Careless responding is prevalent in surveys, but estimates vary on exactly how prevalent CR is in surveys. These estimations range

from 3.5% (Johnson, 2005), 5% (Ehlers et al, 2009), 3-9% (Maniaci & Rogge, 2014), 10-12% (Meade & Craig 2012), 18% (Huang & DeSimone, 2020) and 40% (Osborne & Blanchard) to 60% (Berry et al., 1992). Meade and Poppalardo (2013) found that one CR measure flagged 23% of respondents and that 9 CR measures flagged 8% of respondents as careless (Meade et al., 2017). Furthermore, DeSimone and Harms (2018) found that 51% of participants were flagged by at least one of 5 CR indices whereas only 33% and 14% of participants were flagged by two and three different indices respectively (DeSimone & Harms, 2018). Curran (2016) suggested the modal estimate of prevalence was around 8-12%.

This heterogeneity in the estimates of CR prevalence may result from different studies using different criteria in calculating estimates, the lack of formalization of cutoff usage in CR research, and differences in samples. Johnson (2005)'s estimate of 3.5% was generated from long-string measures, which only capture straight-lining whereas Meade and Craig (2012)'s estimate of 10-12% was based on 17 different measures (Johnson, 2005; Meade & Craig, 2012). Additionally, when dichotomizing measures to classify participants as careless, different studies may be using different cutoffs within measures, which may produce different results. Lastly, different studies have used different types of samples. Ehlers (2009)'s estimate was based on a worker sample whereas other estimates, such as Maniaci and Rogge (2014), were based on student samples (Ehlers et al. 2009; Maniaci & Rogge, 2014).

CR research in low-stakes tests, in which the participants experience no direct consequences as a result of test scores, has suggested that CR is prevalent in tests as well (e.g., Sundre & Wise, 2003; Wise & Kong, 2005). Using self-reported effort and response time effort, Wise and Kong (2005) found response time effort and self-reported effort yielded estimates of 7.5% (n = 37) and 15.8% (n = 75) respectively (Wise & Kong, 2005). Using the same measures and cutoffs for analyses, Rios et al. (2014) found that response time effort and self-reported effort garnered estimates of 14.3% and 23.3% respectively (Rios et al., 2014). Furthermore, using the same cutoffs for response time effort and more lenient cutoffs for self-reported effort, Swerdzewski et al. (2011) found that response time effort and self-reported motivation produced estimates ranging from 11.88% to 22.28% and 35.31% to 44.53% respectively across 6 different tests (Swerdzewski et al., 2011). Lastly, in a review of 20 studies, Rios & Deng found that the mean prevalence of careless responses ranged from 6% to 29% across 5 measurement methods.

Effects of Careless Responding on Data Quality. CR corrupts data quality. Research suggests that CR creates spurious within-group variability that, depending on the scale mean of attentive participants' relationship to the scale midpoint, will inflate or deflate scale scores, which can attenuate (Wise & Kong, 2005, Wise, 2015) or inflate relationships between variables (Huang et al., 2015). Evidence also suggests that CR attenuates effect sizes, statistical power, Cronbach's alpha values and eigenvalues (Huang et al., 2012; Huang et al., 2015; McGrath et al., 2010; Oppenheimer et al., 2009). In

ability tests, careless responding depresses mean scores (Huang & DeSimone, 2020; Rios et al., 2014), reduces the convergent validity of tests (Wise, 2015), and can inflate or deflate relationships between test variables and other variables (Huang & DeSimone, 2020). Additionally, if test-taking effort is unaccounted for, CR can bias item response theory model parameter estimation (Wise & DeMars, 2006) and that an amount of CR that is consistent with prevalence estimates (i.e., 6.5% and 12.5%) can attenuate group level ability estimates by .2 of a standard deviation (Rios et al., 2017).

Research using simulated data has found similar concerning results. Huang et al. (2015) found that prevalence of CR of just 5% and 10% was enough to create spurious relationships between variables (Huang et al., 2015). Woods (2006) found similar confounding effects, in which just as little as 10% of careless responding to reverse-worded items was enough to bias confirmatory factor analysis results. In a recent study, DeSimone et al. (2018) examined the unique effects of different careless response patterns on CR through simulated data (DeSimone et al., 2018). DeSimone et al. (2018) found that random responding attenuates Cronbach's alpha values and principal component analysis (PCA) eigenvalues through decreasing inter-item correlations whereas straight lining inflates Cronbach's alpha values and principal component analysis (PCA) eigenvalues through increasing inter-item correlations (DeSimone et al., 2018). Furthermore, Huang and DeSimone (2020; Study 2) concluded that even small amounts of within-person consistent CR can inflate and attenuate relationships between variables. Furthermore, Wise & Kong, (2005) and Rios & Soland (2021) found in

simulated data evidence that careless responding biases 3 parameter and 2 parameter item response theory parameter estimates.

Data Filtering as a Strategy for Protecting Data Quality. Often researchers will exclude careless participant data from analyses as a means of protecting data quality (i.e., data filtering; Sundre & Wise, 2003; Wise & Kong, 2005). There are some benefits to this approach. Research suggests that even cutting out a small number of careless responders can improve Cronbach's alpha, experimental effect sizes, eigenvalues in exploratory factor analysis and statistical power (Huang et al., 2012; Maniaci & Rogge, 2014; Huang et al., 2012). Using response time effort, research using ability tests has found that filtering careless participants improves mean participant scores (d = .29; Rios et al., 2014) and convergent validity of achievement test scores and the archival Scholastic Aptitude Test scores (change in r of .08-.12 with a median of r = .11; Wise, 2015).

However, in surveys, this exclusion of careless data may come at the cost of external validity. Bowling et al. (2016) found that CR in surveys displays rank-order consistency and is negatively related to informant-rated five factor model personality traits extraversion, agreeableness, conscientiousness, and neuroticism (Bowling et al., 2016). Thus, CR appears to occur systematically within participants. When researchers exclude careless data from analyses, these researchers may be excluding a certain type of people, which may confound or harm the external validity of results.

Given these costs in data and external validity, researchers have investigated methods for preventing CR and factors that potentially affect CR. The latter is of interest because such research aimed at understanding the causes of CR can inform future interventions. In the following sections, I describe survey research concerning a preventative intervention and theorized cause of CR, namely warnings and assessment length respectively. Furthermore, I argue that warnings can effectively prevent CR through raising the stakes of low-stakes tests, and that assessment length will influence CR in tests as well.

Prevention of Careless Responding

Warnings. Researchers have tested the efficacy of different warning manipulations in preventing CR. Some researchers have argued that CR results from insufficient motivation (Meade & Craig, 2012; Bowling et al., 2016). Under this view of CR, warnings then motivate participants through raising the stakes of the assessment by providing a consequence for CR.

The efficacy of these warnings may depend on the consequences researchers warn participants about. To elaborate, the warning tested in Huang et al., (2012) threatened participants with the loss of participation credits if participants responded without sufficient effort. This stern warning was largely successful. The warning induced a significant effect ($p < .05$) on 3 of the 4 CR indices ($d = .25$ to $.43$; Huang et al. 2012). Conversely, the warnings manipulations tested in Meade and Craig (2012) and Ward and

Pond (2015) did not promise the participants that the participants will face direct consequences if they engage in CR. In Meade and Craig (2012), the researchers warned participants that their "responses were subject to [the university's] academic integrity policy" whereas Ward and Pond (2015) simply had informed participants that the researchers would use "sophisticated statistical control methods" to flag CR. Unsurprisingly, neither of these warnings were effective. The warning in Meade and Craig (2012) only had a significant ($p < .05$) effect on 1 of 17 CR indices and the warning in Ward and Pond (2015) only had a significant ($p < .05$) effect on 1 of 6 CR indices. Therefore, the current evidence has suggested that warnings may only be efficacious if they warn the participant of direct and real consequences (e.g., taking away participation credit).

Assessment Length. Often, researchers will administer lengthy item batteries to participants in research studies or validation designs (e.g., Frey & Detterman, 2004; West et al., 2008). As completing assessments is an effortful process (see Sternberg, 1985), researchers have expressed concerns that the use of lengthy assessments may induce fatigue or diminish motivation, which then may induce CR (Meade & Craig, 2012) and affect test performance (Ackerman & Kanfer, 2009), thereby introducing construct irrelevant variance in test scores (Haladyna & Downing, 2004).

Early test research on the topic had validated researcher concerns. Ackerman & Kanfer (2009) report that self-reported fatigue increases as participants progress through

lengthy ability tests (Ackerman & Kanfer, 2009). Furthermore, test research had long observed item positioning effects, in which later item positioning has an upward bias on item difficulty (Meyers et al., 2009). Specifically, Meyers et al. (2009) found that 56% of the variance in changes in Rasch item difficulty across two tests were the result of changes in item positioning, with the same items administered later having greater Rasch item difficulty than those same items administered earlier. Further building off this research, Weirich et al., (2017) found that test-taking effort partially moderated changes in item difficulty, to such that more effortful participants displayed a weaker item position effort compared to less effortful participants.

Survey research found similar results. Galesic (2006) found that participants feel increasingly disinterested and experience greater burden as they progress through a survey (Galesic, 2006). Across several studies in Berry et al. (1992), participants self-report giving more random responses towards the end of a questionnaire than the middle or beginning in several different samples, including a highly motivated sample like job applicants (Berry et al., 1992). Galesic and Bosjnak (2009) found that participants are less likely to complete a longer questionnaire and give less quality responses to questions at the end of a questionnaire (Herzog & Bachman, 1981).

Further extending these investigations to examine the effect of survey length on CR, Gibson and Bowling (2019) found mixed results for the effect of questionnaire length on CR (Gibson & Bowling, 2019). In both studies, participants were randomly

assigned to a short, medium, and long questionnaire (Gibson & Bowling, 2019). Study 1 found a significant effect on two of 4 CR indices, but study 2 failed to replicate these results (Gibson & Bowling, 2019). Bowling et al. (2020) in study 1 and study 2 had participants complete a randomized set of 500 items (Bowling et al., 2020). Questionnaire length displayed a significant effect in 5 of 6 CR measures across two studies (Bowling et al., 2020). Thus, the current evidence suggests that lengthy questionnaires may induce CR in surveys.

Furthermore, for research centered on interventions, researchers need valid measures of CR. In the following section, I discuss the measurement of CR in both surveys and tests. Specifically, from survey research, I discuss the advantages, limitations, and potential use of the infrequency approach, the instructed-response approach, and the long-string approach in capturing CR to tests. Furthermore, I discuss two approaches used in tests and survey research, namely the response time and self-report approaches.

Measurement of Careless Responding.

CR is captured through different approaches. Some approaches involve adding items to the survey or test, whereas other approaches use analyses *post hoc* to capture carelessness. All these approaches capture carelessness indirectly through assumptions. Within this work I will discuss the infrequency, instructed-response, self-report, long-string, response time, and consistency approaches. Thus far, researchers have not entirely

applied survey CR measurement techniques to tests. Specifically, I found no research using the infrequency approach, the instructed-response approach, the consistency approach, and response pattern approach to capture CR in tests. I only found research using the response time and self-report approaches to capture CR. Each of these approaches are limited in ways that other CR approaches can compensate for.

To begin, I first discuss the CR measurement approaches used in both survey and ability test research and practice, namely the self-report approach and the response time approach. Specifically, I define the approach, discuss validity evidence supporting the use of each approach and then discuss the limitations of each approach in capturing CR to tests. Following this analysis, prior to discussing measurement approaches exclusively used in survey research and practice, I discuss the nomological network of an ability test CR measure. Specifically, I describe the ways in which a well-performing CR measure should empirically perform. I then describe the approaches used in ability tests and surveys, namely the response time and self-report approaches. define, examine validity evidence and limitations, and argue for the using the infrequency approach, the instructed-response approach, the consistency approach, and the response pattern approach respectively to capture CR in tests.

Careless Responding Approaches Used in Tests

Self-report Approach: Student Opinion Scale. The student opinion scale captures the self-reported effort expended during a given test and the perceived importance of that test to the participant (Sundre & Moore, 2002). This measure was

developed to account for motivation when conducting research in low-stakes testing situations (Sundre & Moore, 2002). Some examples of items are “Doing well on these tests were important to me” and “I gave my best effort on these tests” (Thelk et al., 2009).

Advantages. The student opinion scale captures both perceived effort put forth in a given assessment and perceived importance of tasks, as opposed to just effort (Sundre & Moore, 2002). The scale has displayed high internal consistency (α = .80-.89; Rios et al., 2014), and has converged with low-stakes test performance (r = .41; Wise, 2006) and a response time-based measure of test careless response (r = .25-.61, median r = .40; Wise, 2015). Furthermore, the scale has consistent factor analytic evidence for its two-factor structure (Sundre & Moore, 2002).

Additionally, research has suggested that the measure performs as it should in regard to test stakes. Under high-stakes testing conditions, in which the results of a test directly affect the participant (e.g., graduate school admissions testing), participants display greater self-reported effort (d = 2.3) and importance of task (d = 1.3) than participants in low-stakes testing conditions (Sundre, 2007). Lastly, unlike other additive measures, researchers have administered and validated the student opinion scale in tens-of-thousands of participants and there is norm data available for the measure (see Sundre, 2007).

Limitations. There are limitations to using the student opinion scale to capture CR. First, the student opinion scale is a global measure. These measures do not tell researchers when and where participants were careless. Additionally, it’s unreasonable to

assume that careless participants will stop and answer the student opinion scale attentively, especially when such a measure is presented at the end of an assessment, where participants are more likely to display careless responding (see Meyers et al., 2009, Weirich et al., 2016). Furthermore, even when careless participants are attentively responding to the self-report scale, it's possible that such participants may engage in faking to disguise their carelessness. Furthemore, Sundre (2007) recommends limiting the use of the student opinion scale to low stakes situations because the scale displays high invariability and, as a result, low internal consistency in high-stakes situations. This result could reflect that under high-stakes conditions participants try harder, but it could also reflect that participants engage in faking and overrate their effort and the importance of the test because of the consequences the test results have upon them. Lastly, for the purposes of data filtering, which describes excluding careless participants from analyses to protect data quality, response time-based CR indices have outperformed the student opinion scale (Rios et al., 2014; Wise & Kong, 2005). To put simply the problems with the student opinion scale, the student opinion scale has questionable validity due to being a self-report scale and does not provide specific information.

Response Time Approach - Page Time and Response Time Effort

The response time approach indexes CR through examining response time data of participants at the page level or item level. The approach assumes that there is a minimum amount of time that a participant has to spend on a given item or page of items to properly process item or page information and respond attentively to that item or page

of items. In survey research and practice, the response time index (called page time in survey research; Huang et al., 2012) is calculated at the page level of analysis whereas the response time index (called response time effort in test research; Wise & Kong, 2005) is calculated at the item level of analysis in test research and practice. Furthermore, given that tests have correct and incorrect answers and are often unidimensional, researchers often use response accuracy and item-total correlation information within response time bands in addition to visually inspecting the response time distribution to set cutoffs at the item level (Wise, 2019), whereas page time uses a static cutoff (Huang et al., 2012). Within this section, I will first briefly discuss the advantages and limitations of using page time in survey research. Following this discussion, I will describe the advantages and limitations of using response time effort in test research and practice.

Page Time Advantages. Using page time to index CR confers advantages. The two second static cutoff is simple and easy to apply when analyzing a dataset. Furthermore, this cutoff was empirically supported in Soland et al., (2019) and Bowling et al., (*2022*). In Soland et al., (2019), the two second cutoff outperformed other cutoffs through identifying a group of participants who displayed greatly higher rates of CR on two other CR metrics than other cutoffs. In Bowling et al., (*2022*), the 2-second cutoff had demonstrated high convergence with a standardized composite index of CR. Page time converges relatively well with other CR indices in recent studies (r = .32 - .67; Ward & Meade, 2018; Bowling et al., 2020; Bowling et al., *2022*), captures the preventative effect of warnings (Huang et al., 2012; Bowling et al., *2022*), and significantly predicts

the ability to recognize study content in participants (r = .74; Bowling et al., *2022*). Furthermore, often researchers are able to collect response time data without alerting participants to the collection of data. Therefore, participant data is unaffected by the collection of data in these contexts. This advantage is not present in other measurement approaches, such as face valid self-report scales, where the participants know what information is being collected and can potentially alter their responses based on that understanding.

Page Time - Limitations. There are some practical limitations in using page time practice and research. First, response time data is only available in assessment modes that allow for the collection of response time data, such as computerized assessments. Therefore, page time cannot be computed in contexts that do not allow for the collection of response time data, such as pen-and-paper assessments or interviews. Furthermore, I found no research concerning what proportion of careless response pages warrants classifying a participant as careless in data filtering. This lack of information on this specific cutoff makes page time difficult to implement optimally in practice and some forms of research, where classifying participants as careless and attentive is needed for data screening.

Response Time Effort - Advantages. There is sufficient evidence supporting the construct validity of response time effort. Concerning the validity of response time effort in tests, response time effort has displayed strong internally consistency, has well researched methods for establishing cutoffs on items (see Wise, 2019), and provides an

item-by-item analysis of carelessness, which allows for the integration of response effort into item response theory models (Wise, 2015). Additionally, research has shown that using response time effort to filter careless participants from test data improves the convergent validity of tests (Wise, 2015). Response time effort has converged with the student opinion scale (r = .25-.61; Wise, 2015) and has not converged with archival cognitive ability test scores, which indicates that it's measuring something distinct from cognitive ability and presumably effort related (Rios et al., 2014; Wise, 2015). Lastly, research has suggested that the participants that response time effort identifies as careless have accuracy rates on items that are close to chance, thereby bolstering the construct validity of the analysis because a test CR index with item-level information should identify a group of responses that exhibit chance level accuracy in the aggregate (DeMars, 2007; Wise & Kong, 2005; Wise, 2015).

Response Time Effort - Limitations. There are some concerning practical limitations in using response time effort. Whereas page time is relatively more flexible in that only response time per page is needed to compute the index, response time effort requires item-level response time information to compute the index. Often, researchers administer tests in settings (e.g., pen-and-paper settings; Weirich et al., 2016) where gathering item-level information is not possible. Furthermore, as Wise and Kong (2005) and others have noted, response time effort may only be an appropriate index of CR in low stakes non-speeded tests because using response time to index CR in a high stakes speeded test may not be able to distinguish between CR and strategic guessing, a

behavior in which participants randomly guess on remaining questions just as time expires to maximize their potential score (Schnipke & Scrams, 1997).

Limitations of Test Careless Responding Measures. Each most common CR indices used in tests is hampered by limitations. Response time effort requires item-level response time information to compute the index. However, item-level response time information is often not possible nor practical to collect (e.g., Weirich et al., 2016). The alternative CR index to response time effort that can be used in such situations, the student opinion scale, has questionable validity. As researchers have noted (e.g., Wise & Kong, 2005), it's unreasonable to assume that careless participants will suddenly provide accurate and attentive information on a self-report scale concerning their own assessment-responding behavior. Going past this faulty reasoning, response time effort has consistently outperformed the student opinion scale through exhibiting larger data filtering effects (Swerdzewski et al., 2011 Wise, 2015; Wise & Kong, 2005) and there is some evidence that suggests that the student opinion scale is susceptible to sources of bias endemic to self-report scales, such as social desirability (see Rios et al., 2014).

Put simply, there is a need for a valid alternative to response time effort that can be used in situations where collection of item-level response time information is not possible. In the following section, I first present a nomological network of a test CR index. Following this presentation, I describe five CR measurement approaches or analyses used in survey research that have not yet been tested in test research, namely the infrequency approach, the instructed-response approach, the consistency approach, and

long-string analysis. I then propose means through which each approach can be applied to capture CR in tests and provide predictions about each index.

Nomological Network

To investigate the construct validity of an index, researchers may develop a nomological network, which describes a predicted pattern of results between the index of interest and several other variables theoretically based on how a well performing measure of a given construct should perform (Bowling et al., 2022; Cronbach & Meehl, 1955; Wise & Kong, 2005). If the results are consistently congruent with the predictions that researchers prior articulated in the nomological network, then researchers can conclude that the evidence supports their interpretation of the measure. Below, I synthesize survey and test CR research to develop a nomological network of a test CR index. This nomological network included several variables and effects, namely the established test CR indices (i.e., response time effort and the student opinion scale), indicators of ability, test performance, inability to recognize item content, response accuracy, the filtering effect, the scale positioning effect, and the warning effect (for a summary of the predictions of the nomological network, see Table 2).

Established Test Careless Responding Measures. The logic for including established CR indices in the nomological network is simple--a measure that researchers interpret to capture a given in tests should converge with previously established indices of that construct (Campbell & Fiske, 1959; Wise & Kong, 2005). Thus, a novel test CR

measure should positively converge with existing test CR indices, namely response time effort and the student opinion scale.

Indicators of Ability. Given that researchers conceptualize CR as an effortful behavior that is unrelated to ability, in construct validation of response time effort and the student opinion scale, researchers have investigated whether these indices were distinct and unrelated to indicators of ability, specifically archival standardized test scores (e.g., Scholastic Aptitude Test scores; SAT) that were obtained in a high-stakes setting (i.e., college admissions). These ability scores need to be from high-stakes testing situations, where participants are responding effortfully because the participants perceive that the test results have direct consequences, because variance in such scores need to reflect differences in ability and not differences in test-taking effort. Results have confirmed the expectations of researchers, exhibiting that archival SAT scores (a proxy for cognitive ability) were weakly correlated with the student opinion scale (r = .14 with SAT-Verbal and r = .01 with SAT-Quantitative; Thelk et al., 2009; Wise & Kong, 2005) and response time effort (range across 9 studies r = -.05 to .19; median r = .08; Wise, 2015). Thus, a novel test CR measure should be uncorrelated or weakly correlated with archival high-stakes standardized test scores.

Test Performance. Building off the discriminant evidence of archival standardized test scores, researchers have further investigated whether CR indices capture effort instead of ability through examining whether such indices are related to test performance and whether the participants that are classified as careless by a given CR

index perform significantly worse on the test than participants that were classified as attentive. Results have demonstrated that test performance is related to the student opinion scale (r = .34 to .58; Rios et al., 2014; Wise & Kong, 2005) and response time effort (r = .54 to .77; Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005). Thus, a test CR index should be related to test performance in addition to being unrelated to minutely related to archival high-stakes ability test scores. If the index is scored such that more CR results in a higher score, then the relationship should be negative, and vice versa.

Inability to Recognize Item Content. Given that careless participants may respond without encoding item information, Bowling et al. (*2022*) argued that careless participants should be less able to recognize item content than attentive participants. The results supported their assertion. Bowling et al., (*2022*) found that item content recognition was significantly related with a composite index of multiple survey CR indices (r = .81). Thus, a CR index should be related to the ability to recognize assessment item content.

Response Accuracy. Given that careless responses are conceptualized as random responses, CR indices should identify a group of responses to items that exhibit chance level accuracy as careless. Conversely, participants that are classified as attentive by such indices should display response accuracy level well above chance. To test an index against this criterion, researchers need item-level information because typically careless

participants are not uniformly careless throughout an assessment. Thus, this criterion has only been studied in response time effort. In empirical tests and through using a variety of cutoff methods, response time effort has been shown to produce response accuracy rates that are close to chance level (Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005; Wise & Ma, 2012; Wise, 2015; Wise, 2019). Therefore, a CR index with item-level information should identify a group of responses to items with chance level accuracy as careless.

Filtering Effect. Researchers have investigated whether using a CR index to classify and exclude careless participants from analyses increases the correlation between the cleaned test scores and an archival measure of ability. Assuming that effort is unrelated to ability, and that the CR index captures effortful responding, then careless responses should introduce construct-irrelevant variance into test scores, thereby attenuating the relationship between the test score and an archival high-stakes ability score. Empirical tests have demonstrated that using the student opinion scale and response time effort to filter data increases the correlation between test scores and archival high-stakes ability scores (DeMars et al., 2007; Thelk et al., 2009; Rios et al., 2014; Wise & Kong, 2005; Wise, 2015; Wise, 2019). Therefore, a CR index should significantly moderate the relationship between archival high-stakes standardized test scores and an external criterion, such that the relationship is attenuated among more careless participants.

Warning Effect. Given that stern warnings have displayed a consistent preventative effect on CR, Bowling et al., (*2022*) have argued that an indicator of a good CR index is whether that index displays a warning effect. Potentially, these results have suggested that stern warnings raise the states of the assessment, thereby promoting effortful responding. Thus, participants exposed to a stern warning should display significantly less CR than participants in a control condition.

Assessment Length Effect. Participants have been shown to engage in CR as they progress throughout a survey (Gibson & Bowling, 2019; Bowling et al., 2020) or test (Weirich et al., 2016). Potentially, this effect may reflect that participants are increasingly fatigued or increasingly lose motivation as they progress throughout a lengthy survey or test. Thus, a test CR index should find a significantly greater amount of CR towards the end of a lengthy test than at the beginning of a lengthy test.

In the following section, I discuss CR approaches used in survey research that has yet to be applied to test CR research. Specifically, I define, and describe the advantages, limitations, and potential use in tests of the infrequency approach, the instructed-response approach, long-string analysis, psychometric synonyms, and the self-report diligence scale. Following this discussion, I use the nomological network above to provide predictions about how each index should perform when applied to capturing CR in tests.

Careless Responding Approaches Used in Surveys

Infrequency Approach. The infrequency approach uses items that to the attentive participant have a clear correct range of responses to detect carelessness (Beach, 1989). These items are also known as bogus items (Meade & Craig, 2012). When using this approach, the researcher is assuming that an attentive participant who reads the infrequency item will respond in the predicted range of responses and that responses that fall outside this range are indicative of carelessness. The content of these items may concern factual impossibilities (e.g. “I was born on February 30th”; Beach, 1989), contain humorous or surprising content (e.g. “All my friends say I would make a great poodle; Meade & Craig, 2012), may inquire about abnormal behavior (e.g. “I eat cement occasionally”; Huang et al., 2014), or may resemble personality items (e.g. “It feels good to be appreciated”; Maniaci & Rogge, 2014). Thus far, research using infrequency has been limited to surveys. I found no research concerning the infrequency approach in ability tests.

Scoring. I observed three different methods for scoring survey infrequency items in the literature. These methods are similar in that each scores one side of a Likert scale as correct, but the methods differ in the size of the range of correct scores or method of scoring responses. In a positively scored infrequency item with a 7-point Likert scale, Huang et al. (2014) dichotomously coded responses 5 (*slightly agree*), 6 *(agree*), and 7 (*strongly agree*) as correct, whereas Meade and Craig (2012) only scored responses 6 *(agree*), and 7 (*strongly agree*) as correct (Huang et al. 2014; Meade and Craig 2012).

The third method from Maniaci and Rogge (2014) continuously scored infrequency items. On a 5-point Likert scale, Maniaci & Rogge (2014) scored the most common response on each item as 0, the response option farthest from the most common response as 4, and the response options in between from 1 to 3. For example, with the item "I don't like getting speeding tickets", the most common response options 5 (*Very True*) was scored as 0, the least common response option 1 (*Not At All True*) was scored as 0, and response options 2 (*A Little True*), 3 (*Some-what True*) and 4 (*Mostly True*) were scored as 1, 2, and 3 respectively (Maniaci & Rogge, 2014).

Advantages. Infrequency items are easy to generate, administer, and score. Huang et al. (2014) demonstrated that infrequency items appear to load onto a single factor and do not induce adverse reactions in participants (Huang et al., 2014). Furthermore, there is evidence suggesting that infrequency items show strong convergence with other CR measures, particularly page time (Bowling et al., 2020), long-string (Bowling et al., 2020; Francavilla, Meade, and Young, 2018), and consistency measures (Meade & Craig, 2012; Francavilla, Meade and Young, 2018). Additionally, this convergence cannot be attributed to common method variance, as infrequency items capture carelessness in a different method than most other CR indices, with the exception of instructed-response items. Lastly, it is likely that infrequency items capture several different careless response patterns. Also, some infrequency items blend in well with common survey items, such as personality items (e.g., "It feels good to be appreciated";

Maniaci & Rogge, 2014). These items may be less distinctive to careless participants and may be better measures of CR.

Limitations. However, there are some limitations in using infrequency items. Infrequency items add to assessment length. Furthermore, infrequency items only detect whether a participant is careless in responding to the infrequency item itself, not other items throughout the questionnaire. Also, infrequency items differ in the number of participants each item classifies as careless (Meade & Craig, 2012; Bowling et al., 2020). Some infrequency items may flag 4.7% of participants as careless (Bowling et al., 2020) whereas others may flag as many as 27% (Meade & Craig, 2012). I found no published research investigating why infrequency items display a heterogeneity in flag rates or what types of infrequency items work best. This heterogeneity cannot be due to different item placements (i.e., items placed later in the questionnaire have larger flag rates due to diminished motivation or increasing fatigue; see section on assessment length) because Bowling et al. (2020) found heterogeneity in flag rates in a survey in which the presentation of items was randomized.

Instead, this heterogeneity may be due to differences in infrequency item content. To elaborate, some items are absurd (e.g., “I am paid biweekly by leprechauns”; Meade & Craig, 2012). These items may flag more participants than others because participants may find it funny to unpredictably agree or disagree with such items. Furthermore, this heterogeneity may be due to perceived ambiguity in items. Curran and Hauser (2019) found that infrequency items sometimes flag attentive participants because participants

misunderstood items. For example, in the item “I sleep less than one hour per night”, a participant justified endorsing this item by reasoning that he or she does sleep less than one hour a night when pulling an “all-nighter”. To the participant, it wasn’t clear whether the item was inquiring about typical sleep behavior (i.e., under conditions in which the participant is not pulling “all-nighter[s]”) or specific conditions (i.e., conditions in which the participant is pulling an “all-nighter”).

Additionally, it’s not clear how many failed infrequency items constitute carelessness. Some researchers have argued for a zero-tolerance approach, in which one failed infrequency item is enough for a researcher to exclude that participant’s data (Osborne & Blanchard, 2011; Periard & Burns, 2014; Kim et al., 2018). However, there are problems with the zero-tolerance approach. A single infrequency item can flag as many as 27% of participants (Meade & Craig, 2012). When using several infrequency items throughout a questionnaire, these items will flag participants that previously were not flagged by other items, thereby increasing the number of excluded participants. This problem is exhibited in study 2 of Kim et al. (2018), in which 67.5% (n = 272) of participants of the first archival sample were identified by at least one infrequency item and 36.4% (n = 265) were identified by at least one infrequency item in the second archival sample (Kim et al., 2018). These large flag rates are concerning because research has pinned the modal prevalence estimate of carelessness in the range of 8-12% (see DeSimone et al., 2015). Thus, to conserve sample size, it may be better to use a more lenient approach. For example, using receiver operator characteristic curves to balance

data elimination and sensitivity to carelessness, Kim et al. (2018) concluded the optimal cutoff for their datasets for exclusion was 4 missed infrequency items out of 9 and 10 infrequency items across two archival samples (Kim et al., 2018). When compared against *Mahalanobis Distance,* a multivariate outlier index, these cutoffs compared favorably to the zero-tolerance cutoff.

Potential Use in Ability Tests. I found no research using infrequency items to capture CR in tests. I propose a method for using the infrequency approach in tests. Typically survey infrequency items present statements with high response invariability among attentive participants to differentiate between attentive and careless participants. The same approach can be used in tests with extremely easy test items. Ideally, these items would be so easy that the item does not differentiate between the ability of participants but rather just differentiates between attentive and inattentive participants. Unlike survey infrequency items, these items would have objectively correct answers and therefore would be easier to score.

For example, the Shipley Institute of Living Scale vocabulary subtest assesses cognitive ability through vocabulary questions (Shipley, 1940). These questions present participants with target words and responses. Participants then choose the word among the responses that is the closest in meaning to the target word. For example, one question presents respondents with the target word “ORIFICE” and responses “brush”, “hole”, “building”, and “lute”, with “hole” as the correct answer (Shipley, 1940). An example of an infrequency item for this test would have the target word “SEAT” with responses

“passion”, “digress”, “chair”, and “exhibition”. This item not only presents a simple synonym pair, but also response options that are completely unrelated to the pair itself. Therefore, these questions should display high response invariability with high accuracy rates among attentive participants, whereas careless participants should exhibit higher variability among these items with chance-level accuracy rates. Put differently, attentive participants should have no problem getting this question correct whereas careless participants may not get this problem correct.

Ideally, following the criteria set by Wise and Kong (2005), a well performing test careless responding index will converge with other CR indices, display internal consistency (assuming the metric is appropriate the index), classify participants that perform worse and whose data attenuates scale validity as careless (i.e., displays filtering effects), classify participant responses that display on average rates of getting items correct that are consistent with random chance. Furthermore, a test CR index should just capture whether or not the participant is responding with regard to item content instead of ability. Thus, a well-performing test careless responding index should display discriminant validity through not correlating or correlating very weakly with academic ability (see Wise & Kong, 2005).

Under my nomological network, if infrequency items capture CR, infrequency items should demonstrate several results. Specifically, infrequency items should positively converge with established test CR indices (i.e., student opinion scale and response time effort), converge with test performance, exhibit filtering effects with

archival standardized test scores and grade point average, exhibit warning effects, and exhibit assessment scale positioning effects.

Hypothesis (1a): Infrequency items will positively and significantly converge with the student opinion scale and response time effort.

Hypothesis (1b): Using Infrequency items to filter careless participants will improve the convergent validity between cognitive ability tests and archival standardized test scores.

Hypothesis (1c): Using infrequency items to filter careless participants will improve the correlation between cognitive ability test scores and undergraduate grade point average.

Hypothesis (1d): Infrequency items will significantly and positively correlate with test performance.

Hypothesis (1e): Infrequency items will not be correlated with archival standardized test scores.

Hypothesis (1f): Participants assigned to the warning condition will significantly display less CR on the infrequency index than those participants in the control condition.

Hypothesis (1g): Participants who take a target battery of items at the end of the assessment will significantly display more CR on the infrequency index than those participants who take the target battery at the beginning of the assessment.

Instructed-Response Approach

The instructed-response approach uses items with clear and direct instructions on responding behavior to determine whether a participant is careless (DeSimone et al., 2015). In using this approach, researchers assume that an attentive participant will correctly comply with the instructions embedded within the item. Instructed-response items are also sometimes called trap questions (Liu & Wronski, 2018). These items may state the directions directly (e.g., "Please select strongly disagree for this item") and or embed the directions within the facade of a typical item (e.g.,"I am competent in panabogy-skip this item to show that you have read survey items correctly"; Kam & Chan, 2018; see Liu & Wronski, 2018 for more elaborate examples). Scoring instructed-response items is simple. Researchers score participant responses that are the instructed response as attentive and any response that deviates from the instructed-response as careless (Curran, 2016).

Advantages. Using instructed-response items to capture carelessness carries advantages. Instructed-response items are easy to generate, administer, and score. Instructed-response items converge well with other measures of CR, such as consistency measures (r = .32 - .55; Francavilla, Meade & Young, 2017; Kam & Cham, 2018; Ward & Meade, 2018) and page time (r = .47; Ward & Meade, 2018). The instructed-response item format is easily manipulated and can readily fit different contexts.

Limitations. There are some concerns and limitations in using instructed-response items. Like infrequency items, instructed-response items add to survey or test length, only tell whether a participant is careless in responding to the instructed-response

items themselves, and display heterogeneous flag rates. To elaborate on this last point, some items flag 3.8% of participants (Kam & Chan, 2018) whereas others flag 27% of participants (Liu & Wronski, 2018).I found no research investigating the cause of this heterogeneity.

This heterogeneity may be due to item differences. For example, it may be that this item "I like people in general and please skip this item…" (23% flagged as careless) flagged more participants as careless than "Select strongly agree for this item" (14% flagged as careless) because the former item embedded the instructions within a facade of a typical item (Kam & Chan, 2018). This facade may make the items more sensitive because participants are reading and responding to the facade as opposed to reading further and responding to the instructions. More research is needed to determine the cause of this heterogeneity.

Lastly, the current research is also not clear as to how many instructed-response items should warrant exclusion or participant data or how many instructed-response items should be included within a survey or test. Kam and Chan (2018) demonstrated that stricter cutoffs further improve the negative correlation between positively scored and reverse scored items (Kam & Chan, 2018). More research is needed to definitively answer this question.

Potential Use in Tests

I found no research extending instructed-response items to capturing carelessness in ability tests, but in the same way that instructed-response items are adapted to fit the

format of survey items (e.g. Maniaci & Rogge, 2014), I propose that instructed-response items can be embedded easily and successfully within verbal tests (e.g., Shipley, 1940). I condition my statement to verbal tests because instructed-response items rely on words to instruct the participants in their responding behavior.

To provide an example of an instructed-response item that blends with the test format, the Shipley (1940) abstraction subtest assesses cognitive ability through presenting an incomplete series of symbols, letters, or numbers to the participant. Then, through open-response, participants complete the series. An example is “mist is wasp as pint in tone --” with “on” as the correct response. An example of an instructed-response question that blends within this format would be “please enter carrots --”. Just as with infrequency items, attentive participants should have no problem complying with item instructions and getting the item correct whereas careless participants will likely not get the item correct. Therefore, these items should differentiate between careless and attentive participants.

Under my nomological network, if instructed-response items capture CR, instructed-response items should demonstrate several results. Specifically, instructed-response items should positively converge with established test CR indices (student opinion scale and response time effort), converge with test performance, exhibit filtering effects with archival standardized test scores and grade point average, exhibit warning effects, and exhibit assessment scale positioning effects.

Hypothesis (2a): Instructed-response items will positively and significantly converge with the student opinion scale and response time effort.

Hypothesis (2b): Using instructed-response items to filter careless participants will improve the convergent validity between cognitive ability tests and archival standardized test scores.

Hypothesis (2c):Using instructed-response items to filter careless participants will improve the correlation between cognitive ability test scores and undergraduate grade point average.

Hypothesis (2d): instructed-response items will significantly and positively correlate with test performance.

Hypothesis (2e): instructed-response items will not be correlated with archival standardized test scores.

Hypothesis (2f): Participants assigned to the warning condition will significantly display less CR on the instructed-response items than those participants in the control condition.

Hypothesis (2g): Participants who take a target battery of items at the end of the assessment will significantly display more CR on instructed-response items than those participants who take the target battery at the beginning of the assessment.

Consistency Approach

The consistency approach captures CR through examining the within-person consistency of responses to items. Given that random responding is random, a response

to one item should not give information on how that participant will respond to another item. Using this logic, the consistency approach assumes that attentive responding should be predictable and therefore consistent across similar items or constructs and inconsistent across dissimilar items or constructs (Curran, 2015). Therefore, under this approach, inconsistent responses to highly correlated or parallel items and consistent responses to highly diverging items reflect carelessness.

Psychometric Synonyms and Antonyms. The most popular index of the consistency approach is psychometric synonyms and antonyms, in which researchers use highly correlated pairs of items to identify careless respondents (Curran, 2015). In psychometric synonyms, highly positively correlated pairs are used whereas highly negatively correlated pairs are used in antonyms. These pairs are identified after data is collected. I observed two methods for identifying pairs. The first involves using a predetermined cutoff (i.e., $r = .6$) to select pairs. Pairs above this correlation cutoff are included in the analysis (Curran, 2015). Another method simply just takes the top most correlated pairs from the dataset. For example, Huang et al. (2012) had selected the 30 most negatively correlated item pairs for their computation of psychometric antonyms. I found no published research validating the convention cutoff ($r = .6$) or the approach used in Huang et al. (2012). These methods have been used in conjunction (see Johnson, 2005). Concerning classification cutoffs, Huang et al. (2012) placed cut score at a point where the correlation is low (e.g., 0, -.03) or in the opposite direction (e.g., 0.22 for psychometric antonyms; Huang et al., 2012).

Advantages. There are several notable advantages to the psychometric synonyms and antonyms index. Firstly, unlike other methods (e.g., infrequency), the analysis does not require adding specific items to the assessment. Thus, the analysis can be applied to any dataset with a sufficient number of item pairs. Furthermore, psychometric synonyms and antonyms display moderate to strong convergence with other CR indices, such as page time (r = .66), instructed-response (r = .48-.54), and infrequency items (r = .37 - .66; Huang et al., 2012; Meade & Craig, 2012), thereby indicating that antonyms and synonyms are capturing a construct that is similar to that which is being captured by other indices.

Limitations. There are some gaps in knowledge or practical hurdles that limit the application of psychometric synonyms and antonyms. First, I found no research determining how many pairs of items are needed or how to empirically determine classification cutoffs within any individual datasets. Furthermore, even in a lengthy assessment, using the conventional cutoff (r = .6) net a concerningly low amount of item pairs. For example, Meade and Craig (2012) only found 5 item pairs for their psychometric antonym index in a survey with over 500 items. Of course, whether 5 item pairs are sufficient for using psychometric antonyms hasn't been empirically determined. However, this sparsity of pairs presents a practical problem. Given that Meade and Craig (2012) only found 5 pairs that surpassed the conventional cutoff in such a large survey, it's conceivable that researchers may have trouble finding any pairs in shorter surveys.

Potential Use in Tests. Just as it is the case in surveys, careless responses in tests should provide little information to how a respondent will respond on other items. Substantiating this claim, Wise (2019) found that careless responses have item-total correlations Thus, examining the correlation between highly positively correlated pairs of test items may be useful in identifying careless respondents.

However, antonyms may likely not be useful in capturing CR in tests because highly negatively correlated item pairs are unlikely to occur in tests. In surveys, scales often have reverse scored items or scales of different constructs that would naturally produce heavily correlated item pairs. Conversely, tests do not have an equivalent to reverse scored items and, typically, batteries of ability tests do not assess heavily divergent constructs. Thus, I expect synonyms to only be practically applicable to capturing CR in ability tests.

Under my nomological network, if the psychometric synonyms index captures CR, psychometric synonyms should demonstrate several results. Specifically, psychometric synonyms should positively converge with established test CR indices (i.e., student opinion scale and response time effort), converge with test performance, exhibit filtering effects with archival standardized test scores and grade point average, exhibit warning effects, and exhibit assessment scale positioning effects.

Hypothesis (3a): The psychometric synonyms index will positively and significantly converge with the student opinion scale and response time effort.

Hypothesis (3b): Using the psychometric synonyms index to filter careless participants will improve the convergent validity between cognitive ability tests and archival standardized test scores.

Hypothesis (3c): Using the psychometric synonyms index to filter careless participants will improve the correlation between cognitive ability test scores and undergraduate grade point average.

Hypothesis (3d): The psychometric synonyms index will significantly and positively correlate with test performance.

Hypothesis (3e): The psychometric synonyms index will not be correlated with archival standardized test scores.

Hypothesis (3f): Participants assigned to the warning condition will significantly display less CR on the psychometric synonyms index than those participants in the control condition.

Hypothesis (3g): Participants who take a target battery of items at the end of the assessment will significantly display more CR on the psychometric synonyms index than those participants who take the target battery at the beginning of the assessment.

Self-Report Approach

Self-Reported Diligence. The self-report diligence scale assesses how attentive and conscientious participants think they were in responding to items (Meade & Craig, 2012). In using this measure and other self-report measures, the researcher is assuming that careless participants, who have historically responded to items without regard to item

content, will respond attentively to these items. This scale originated in Meade and Craig (2012), in which the authors introduced 17 items to assess self-report scales to assess participant attentiveness and engagement (Meade & Craig, 2012). An exploratory factor analysis on attentive respondent data revealed that the 9 of the 17 items loaded onto one factor (Meade & Craig, 2012). Examples of these items are "I carefully read every survey item", "I put forth my best effort in responding to this survey" and "I was actively involved in this study" (Meade & Craig, 2012). These items are presented on a 7-point Likert scale and typically larger mean score values represent more carelessness (Meade & Craig, 2012; Ward & Meade, 2018).

Advantages. The self-reported diligence scale is internally consistent (α = .83-91; Meade & Craig, 2012; Ward et al., 2017) is easy to administer within a survey or test, and modestly converges with other CR indices (r = .18 - .51; Meade & Craig, 2012). Additionally, in logistic regression diligence performed well at correctly identifying participants as careless or attentive and accounted for more pseudo R^2 variance than other CR indices, such psychometric antonyms, Long-String, total response time, and Mahalanobis *D* (Meade & Craig, 2012).

Limitations. There are limitations and disadvantages in using the diligence scale. The first concern is regarding the assumption in using self-report indices to capture carelessness. As some researchers have noted (see Wise, 2015), if participants are careless throughout a questionnaire, it is unreasonable to assume that those participants will stop and respond attentively to the diligence scale. Additionally, as Wise (2015)

notes, CR scales like the self-report diligence scale are global measures, these scales do not tell researchers where exactly participants were careless, nor do they tell researchers about the nature of their careless behavior (Wise, 2015). To elaborate on this latter point, a low response score to the diligence item "I could've paid closer attention to the items than I did" (Meade & Craig, 2012) does not tell whether the participant could've paid better attention to all items or some items in particular. The information the scale provides is ambiguous. Lastly, whereas the diligence scale does converge with other CR indices, this convergence is weaker than the convergence seen in other indices, and evidence has suggested that diligence loads onto a separate factor consisting solely of self-report CR indices (Meade & Craig, 2012). This latter point could suggest that self-report indices capture a separate dimension of carelessness or that self-report indices are capturing a dimension that is separate but related to carelessness. No further research was found exploring the factor structure of multiple CR measures. Furthermore, the discriminant validity of the self-report diligence scale is unclear because I found no research investigating whether diligence scale scores were uncontaminated by known corrupters self-report scale scores, such as impression management and self-deception.

Potential Use in Tests. I have identified no research using the self-report diligence scale in tests. To administer this measure post-test, some items would need to be modified because the items are contextualized to surveys (e.g. "I carefully read every read survey item"). However, I propose that after these modifications researchers can use the self-reported diligence test to capture carelessness in tests.

Under my nomological network, if the self-report diligence scale captures CR, self-report diligence should demonstrate several results. Specifically, psychometric synonyms should positively converge with established test CR indices (i.e., student opinion scale and response time effort), converge with test performance, exhibit filtering effects with archival standardized test scores and grade point average, exhibit warning effects, and exhibit scale positioning effects.

Hypothesis (4a): The self-report diligence scale will positively and significantly converge with the student opinion scale and response time effort.

Hypothesis (4b): Using the self-report diligence scale to filter careless participants will improve the convergent validity between cognitive ability tests and archival standardized test scores.

Hypothesis (4c):Using the self-report diligence scale to filter careless participants will improve the correlation between cognitive ability test scores and undergraduate grade point average.

Hypothesis (4d): The self-report diligence scale will significantly and positively correlate with test performance.

Hypothesis (4e): The self-report diligence scale will not be correlated with archival standardized test scores.

Hypothesis (4f): Participants assigned to the warning condition will significantly display less CR on the self-report diligence scale than those participants in the control condition.

Hypothesis (4g): Participants who take a target battery of items at the end of the assessment will significantly display more CR on the self-report diligence scale than those participants who take the target battery at the beginning of the assessment.

Long-String Analysis. Long-string analyses are performed *post hoc* on data to determine whether participants are repeatedly inputting the same response, thereby forming a long-string of response inputs (Curran, 2015; Johnson, 2005). The analyses assume that CR is reflected in consecutive responses of the same input that surpass a predetermined length cutoff (Johnson, 2005; Curran, 2016). DeSimone et al. (2015) recommended using long-string analyses in surveys with measures of multiple dimensions or across items that are positively and reverse scored.

I observed two kinds of long-string analyses, maximum long-string and long-string average . A maximum long-string analysis yields the longest string of inputs for each participant per page (Johnson, 2005; Meade & Craig, 2012; Ward & Pond, 2015). If a participant produces a maximum string of responses that passes a cutoff on a given page, then that participant receives a 1 (careless) instead of a 0 (attentive) for that page (Johnson, 2005; Meade & Craig, 2012; Ward & Pond, 2015). If the summed maximum values for each participant passes a cutoff, then the participant's data is regarded as careless data (Johnson, 2005; Meade & Craig, 2012; Ward & Pond, 2015). Other studies have conducted a long-string average analysis to compare the average maximum string of across pages to a cutoff (Meade & Craig, 2012; Francavilla, Meade & Young, 2018; Gibson & Bowling, 2019).

Cutoffs. In contexts in which classifying participants as careless and attentive is necessary (i.e., excluding careless participants from data to protect analyses), researchers have to set cutoffs. I found no research systematically comparing different methods for generating cutoffs, but researchers appear to have different ways of establishing cutoffs. For example, Huang et al. (2012) measured maximum strings and established cutoffs for each of the 5 response options, such that each participant had a long-string maximum long-string value for response option 1 (*Strongly Disagree*) that was compared to the established cutoff for response option 1, and so on (Huang et al., 2012). Conversely, Francavilla, Meade and Young (2018) established a cutoff at the next rounded whole number maximum long-string value that was two standard deviations from the maximum long-string value mean (Francavilla, Meade & Young 2018).

Advantages. Long-string analyses have some advantages. Long-string analyses are simple to compute, do not require adding items to a survey, show good convergence with infrequency items and page time in recent studies (r = .39 - .66; Gibson & Bowling, 2019, Bowling et al. 2020) and can be performed on any dataset in which a careless string of responses may reasonably occur.

Limitations. There are some concerns over using long-string analyses to capture carelessness. Barring some exceptions (Costa & McCrae, 2008), cutoffs in the literature for long-string analyses are not established. Researchers have seemed to establish cutoffs for each study after data collection. Whereas convergence in recent studies is promising,

the convergence of long-string with other CR indices has historically been low (r = .09 - .3; Huang et al., 2012 Meade & Craig, 2012; Francavilla, Meade & Young, 2018). In Meade and Craig's (2012) factor analysis, the long-string analyses loaded onto a factor separate from all of the other measures (Meade & Craig, 2012). This loading may indicate that long-strings capture only straight-lining, which is a careless response pattern of identical-consecutive inputs (Schonlau & Toepoel, 2015), whereas other measures such as infrequency items capture both straight-lining and random responding, which is a response pattern in which careless respondents imbue their responses with spurious variance, possibly to disguise carelessness (DeSimone et al., 2018). However, depending on the context, this limitation may be an advantage in contexts where capturing specifically straight-lining is necessary, such as research investigating the effects of different response patterns (e.g., DeSimone et al., 2018). Lastly, on a scale with few or no reverse worded items, conceivably, researchers can classify an attentive participant with an extreme score on a construct as careless (DeSimone et al., 2015).

Potential Use in Ability Tests. I found no research using long-string analyses in tests. However, I propose that any test that utilizes response options that can utilize long-string analyses. Given that tests vary in their correct response options, participants who respond in strings likely are exhibiting carelessness just as participants would be if those participants responded in a string on a list of survey items. However, I propose that long-strings are potentially more useful in tests than surveys. As DeSimone and Harms (2015)

noted, long-string analyses would falsely flag attentive participants who have extreme values on a trait who are responding to a survey scale with few or no reverse-worded items. Conversely, this false positive cannot happen in tests because test items are not composed of rating scales that are anchored by extreme levels on a trait (e.g., *strongly agree* to *strongly disagree*, *bad* to *good*; Dane 1990). Therefore, in tests that vary in the correct response option, long-strings in tests should be more effective.

Under my nomological network, if long-strings capture CR, long-strings should demonstrate several results. Specifically, psychometric synonyms should positively converge with established test CR indices (i.e., student opinion scale and response time effort), converge with test performance, exhibit filtering effects with archival standardized test scores and grade point average, exhibit warning effects, and exhibit scale positioning effects.

Hypothesis (5a): Long-strings will positively and significantly converge with the student opinion scale and response time effort.

Hypothesis (5b): Using the long-strings to filter careless participants will improve the convergent validity between cognitive ability tests and archival standardized test scores.

Hypothesis (5c):Using long-strings to filter careless participants will improve the correlation between cognitive ability test scores and undergraduate grade point average.

Hypothesis (5d): Long-strings will significantly and positively correlate with test performance.

Hypothesis (5e): Long-strings will not be correlated with archival standardized test scores.

Hypothesis (5f): Participants assigned to the warning condition will significantly display less CR on the long-string index than those participants in the control condition.

Hypothesis (5g): Participants who take a target battery of items at the end of the assessment will significantly display more CR on the long-string index than those participants who take the target battery at the beginning of the assessment.

Factor Structure of Careless Responding Indices.

There is limited research on the latent factor structure of CR indices. I identified 1 factor analysis on CR indices in the literature. In Meade & Craig (2012), the authors performed an exploratory factor analysis on 17 different CR indices to investigate the factor structure of these indices. The authors found that a three-factor structure best fit their CR data. The first factor contained indices that conceptually capture a variety of response patterns, such as random responding and straight-lining (see Long-String limitations). The second factor contained self-report indices, thereby indicating that self-report carelessness is related, but different from carelessness as measured by other indices. Lastly, the two long-string analyses loaded onto the third factor, possibly because long-string analyses only measure straight-lining.

I posited that the same three-factor structure observed in research on CR in surveys would best capture the variance in CR used in tests. Furthermore, the Meade & Craig (2012) exploratory factor analysis did not contain response time based measures of CR. Given that page time would capture both random responding and straight-lining, a response time based index should load onto the same factor as other CR indices in tests that also capture random responding and straight-lining, such infrequency items, psychometric synonyms, and instructed-response items. Conversely, given that long-string analyses only capture straight-lining, these analyses should load onto a separate factor. Lastly, given that the student opinion scale and self-report diligence both capture self-reported CR, they should load onto the same factor as well in tests.

Hypothesis (6): A three-factor structure with careless responding indices will demonstrate the best fit when compared to other models, with infrequency items, psychometric synonyms, instructed-response items, and response time effort loading onto one, self-report diligence and the student opinion scale loading onto a second factor, and the long-string analysis loading onto a third factor.

II. METHOD

Descriptive Statistics. 291 students enrolled at a Midwestern University participated online in my study. I aimed to collect 278 participants, which I determined using GPower would have given 80% power to detect an effect size (Cohen's *d*) of .3. By the time I had noticed the participant count passed that number, data had been collected on 291 participants, and I had decided to include those who had participated after the 278th participant. I gave course credit to participants in return for their participation. The average reported age of the 291 participants was 20.99 and 70.1% of participants reported that they were female. Given that few participants reported SAT scores (*n* =69), I dropped self-reported SAT scores from analyses. The average self-reported GPA was 3.01 (SD = .80), and the average self-reported ACT score was 21 (SD = 5.98). Additionally, 69.420% of participants were Caucasian, 18.21% were African-American, 3.43% were Hispanic, 4.81% were Asian-American, .03% were American Indian or Alaskan Native, .03% were Pacific Islander, and 3.43% were Middle Easterner.

Design

My study's design is a 2 x 2 between-subjects design. I randomly assigned participants to either a warning condition or a control condition. Additionally, I randomly assigned participants either to a condition with a target assessment at the beginning of the test or assigned participants to a condition with a target assessment at the end of the test. I further describe these manipulations below. Following this description, I describe my

dependent variables, which are CR indices, namely response time effort, the student opinion scale, the infrequency index, psychometric synonyms, long-string, the self-report diligence scale, and the instructed-response index.

Manipulations

Warning Manipulation. I randomly assigned participants to two conditions: a warning condition or a control condition. In the warning condition, I used the warning message from Huang et al. (2012) to inform participants at the beginning of the test that I would use advanced statistical methods to assess whether or not they responded carefully to test items. Furthermore, I then told the warning condition participants that I would revoke participation credits had these analyses determined their response behavior to be careless. Furthermore, I used the 2 manipulation check items from Huang et al. (2012) to assess the effectiveness of my warning manipulation. These items are "The researcher told me that he or she will use advanced statistical techniques to detect the accuracy and thoughtfulness of my responses" and "that I will lose my research credits if I fail to provide accurate and thoughtful responses to today's survey questions.". I administered each item to participants on a 7-point graphic rating scale from 1 (*strongly disagree*) to 7 (*strongly agree*) after the tests at the end of the experiment.

Scale Positioning Manipulation. I used three types of measures in my study. I used a target assessment, filler test measures, and CR indices. I used the Shipley Institute of Living Scale (Shipley, 1940) as my target assessment. I described this scale in the

following section. I manipulated the position of the target assessment to capture the effect of length on CR. To elaborate, if length facilitates CR, then participants who receive the target assessment at the end of the test should display greater rates of carelessness in responding to that scale. Filler test measures were included for three purposes. The first purpose was to fill out space for the scale positioning manipulation. Secondly, Included test items that capture constructs and criteria to test hypotheses concerning data filtering. The third purpose of these filler items was to have items that I can embed careless indices in. Lastly, I included the CR indices to capture carelessness.

Measures

I describe each specific measure that I used below. I used the Shipley Institute of Living Scale (Shipley, 1940) as my target assessment. For the filler test items and target assessment, I describe each measure, cite validity data, and how I used each measure when necessary. For CR indices, I describe the measures and cutoffs used.

Target Assessment. I used the 60 item Shipley Institute of Living Scale as the target assessment (SILS; Shipley, 1940). Shipley (1940) originally developed the scale to capture intellectual impairment, but researchers increasingly over time have used the SILS as a non-proprietary and brief measure of intelligence (Weiss & Schell, 1991). The SILS contains a 40 item vocabulary subtest that assesses verbal ability and a 20 item abstraction subscale that assesses inductive reasoning. Participants have 10 minutes to complete each subtest.

The vocabulary subtest presents participants with a word as the item stem and 4 words as response options. Participants then identify which word among the response options is closest in meaning to the item stem word. An example item has the item stem "TALK" with the response options as "draw", "eat", "speak", and "sleep", with "speak" as the correct answer (Shipley, 1940). The abstraction subtest presents participants with an uncompleted series that the participant then completes. An example item is "1 2 3 4 5 - ", with 6 as the correct answer (Shipley, 1940). I will score a correct response as 1, and an incorrect response is scored as 0. I will then combine points across items into a total score.

The SILS demonstrates good validity. Both subtests of the SILS are internally consistent (.87-.92; Shipley, 1940) and the SILS displays test-retest reliability (r = .80; Martin et al., 1977). The SILS converges well with other measures of intelligence, such as Raven's progressive matrices (r = .57-.72; Eisenthal & Hartford, 1971), the verbal subtest of the Kaufman Brief Intelligence Test (r = .59-81; Bowers et al, 1998), the Slosson Intelligence Test (r = .54; Martin et al., 1977), and the Wechsler Adult Intelligence Scale (r = .78-.86; Weiss & Schell, 1991).

Filler Test Measures - Miller Analogies Test Practice Items. I used a sample of 80 items from 501 practice items for the Miller Analogies Test (LearningExpress, 2002). The Miller Analogies Test uses verbal analogies to assess cognitive ability. Since the test's development in 1926, schools and researchers have used the Miller Analogies Test

for the purposes of selection (Kuncel et al., 2004; Meagher et al., 2021). Item stems consist of an incomplete analogy and participants then select from 1 of 4 response options to complete the analogy. An example item is "_____ : trail :: grain : grail" (LearningExpress, 2002). This analogy is understood as _____ is to trail as grain is to grail. Among the response options "train", "path', "wheat", and "holy", train is the correct answer (LearningExpress, 2002). Practitioners typically administer the test with 120 items and give participants 60 minutes (30-seconds per item) to complete the test (Meagher et al., 2021). I will score a correct response as 1, and an incorrect response is scored as 0. I will then combine points across items into a total score.

The Miller Analogies Test displays internal consistency, convergent validity, and predictive validity. Using meta-analysis, Kuncel et al. (2004) found that the Miller Analogies Test correlated strongly with other measures of cognitive ability, such as the GRE (ρ = .57 - .88). Additionally, Kuncel et al. (2004) also found that the Miller Analogies Test predicted 1st-year graduate grade point average (ρ = .41), overall graduate grade point average (ρ = .39), and faculty ratings (ρ = .37).

Careless Responding Indices. I included seven measures of CR or related constructs in my study to both capture carelessness and to investigate my hypotheses and research questions. I included: (1) infrequency measures, (2) instructed-response measures, (3) the self-reported diligence scale, (4) the student opinion scale, (5) psychometric synonyms, (6) long-string analyses, and (7) response time effort. I provided

validity information for each of these measures earlier in the introduction. Below, I describe how I applied each measure or analysis in my study.

***Infrequency Items*.** I created two sets of infrequency items to fit the two test measures I used in my study, namely the set of practice Miller Analogies Test items and the Shipley Institute of Living Scale. Below I describe the process I used to design items for the set of practice Miller Analogies Test items and the Shipley Institute of Living Scale.

Miller Analogies Test Practice Items. I created a set of 10 Miller Analogies Test infrequency items. I created these items with the intention of having extremely little response variance among attentive participants. Specifically, I intended to create analogies that ideally any attentive participant can intuitively solve from looking at the item stem alone without consulting the response options. An example of an item stem is "kitten : cat :: puppy : _____" (for the full set of items, see Appendix A). Furthermore, in the response options, I surrounded the correct answer with unrelated words that were generated by a random word generator. I did this to further ensure that any attentive participant could get these items correct. In the aforementioned example, the response options are "dog" (correct response), "commerce", "serious", and "straw". For the instructed-response items, I took practice items and altered the stem to refer to one response option. For example, I took an item with the response options of "armor",

"belt", "tyne", and "shoe" and altered the stem to "_____ : please :: select : armor" (for a full list of items, see appendix). I included 4 of these instructed response items.

Shipley Institute of Living Scale. I created 7 infrequency items for the SILS vocabulary and abstraction subscales using the same approach I used for the Miller Analogies Test. I intended to create items that ideally any participant would get correct. An example of a vocabulary stem is "baby". Response options for this example are "child" (correct answer), "article", "army", and "productive". An example of an uncompleted infrequency abstraction subscale series is "A B C D -". I included 3 of these items. For the complete list of items, see Appendix x. I will score correct responses as attentive (0) and incorrect responses as careless (1). I will then create sums for each subscale and for the total scale.

For the SILS, I also created instructed response items. An example of instructed-response items for the SILS vocabulary subtest is "please select rain" with "lemon", "Rain", "Like", "Heal", and "Gloom" as the response options (for the complete list of items, see Appendix). An example of an instructed response item for the SILS-A is "please enter 2 below -". I included 2 of these items. I will score the instructed response as 0, and all other responses as 1. I will then combine points across items into a total score.

Psychometric Synonyms. For psychometric synonyms, I will run exploratory analyses to determine the adequate item-pair correlation cutoff for synonym pairs. In

survey research, typically researchers have taken item-pairs that exceed $r > .6$ (Curran, 2016). However, this convention may not be appropriate for tests. Thus, I will test a set of cutoffs (r = .55, .6, .65, .7, and .75 respectively) against a criterion (i.e., convergence with other CR indices) to determine the appropriate cutoff. Once I have the cutoff for item inclusion, I will then compute the correlation between pairs. A higher positive correlation reflects that the participant was more attentive.

Self-Reported Diligence Scale. I administered the Meade & Craig (2012) self-reported diligence scale to participants at the end of the study. I modified some items because any of the items of the scale are worded for surveys (e.g., "I carefully read every survey item"; Meade & Craig, 2012). I administered response options on a graphic rating scale that ranged from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). A full list of items is provided in appendix x. A full list of these items are included in Appendix x. I will reverse score items 1, 4, 5, and 8 and I will positively score items 2, 3, 6, 7, and 9. I will then sum these scores. A higher score will reflect greater carelessness in taking the assessment.

***Student Opinion Scale*.** I administered the student opinion scale to participants at the end of the study. I counterbalanced the order of the student opinion scale and the self-reported diligence scale to control for order effects. I administered response options on a graphic rating scale that ranged from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). A full list of these items are included in Appendix x. I will reverse score items 1, 2, 5, 6, 8, and

10 whereas I will positively score items 3, 4, 7, and 9. A higher score will reflect greater carelessness and that the assessment was not important to the participant. Some examples of items are “Doing well on these tests were important to me” and “I gave my best effort on these tests” (Thelk et al., 2009)

Response Time Effort. I collected response time data for each participant at the item level for tests. I will run a series of exploratory analyses to determine which cutoff technique is appropriate for this data collection. I will first attempt to use the modified visual inspection with conditional response accuracy described in Wise (2019). In assuming that response time data is bimodal due to overlap between distributions of careless and attentive participants and that careless participants should have accuracy rates near chance, this method uses visual inspection of response time and accuracy data to set response time cutoffs at the item level.

If this method is not usable due to significant temporal overlap between careless and attentive participants, I will use the change in information method introduced in Wise (2019). The change in information method utilizes item-total correlations to determine response time cutoffs (Wise, 2019). This method assumes that careless responses are uninformative and provide low item-total correlations. Thus, response data in time bands filled primarily with careless participants should yield low item-total correlations. Research has supported both of these methods (Wise, 2019) If this method is inadequate, I will use the normative threshold (see Wise & Ma, 2012; Rios et al., 2014) method. This

method uses a percentage of the mean response time per item as a cutoff. Once I have determined the appropriate cutoff method, I will score a response time that is quicker than the cutoff as careless (0) and a response time that exceeds the cutoff as attentive (1). I will then compute a summed average of the scores.

Procedure. I recruited participants through an online university medium for research recruitment. Participants obtained a link to the study from this website. Participants then completed the study in environments of their choice without face-to-face contact with the researcher. Furthermore, I presented assessments to participants in blocks (see Table 3). There are 5 blocks, the first which contains the warning message, the second which consists of the target assessment (i.e., the SILS), the third which has the filler measures, and the forth which consists of the item-content recognition, self-report diligence, and student opinion scale CR indices, and the fifth which contains the warning manipulation check item. I randomized the presentation of different measures within blocks. For example, for the CR indices block (i.e., block 5), each participant responded to the self-report diligence scale, and the student opinion scale in a random order. Within each scale, to avoid the confounding effect of strategic rapid-guessing for response time effort (see Schnipke & Scrams, 1997) and to obtain item-level information, I will present each test item individually. Survey items will be presented normally on a page.

The specific condition determined the order of presentation of blocks. In the condition with no warning and with the target assessment at the beginning of the

assessment, I administered participants first the target assessment, then the filler measures, the CR indices, and lastly the warning manipulation check items. For the condition with no warning and the targettarget assessment at the end of the assessment, I administered participants first the filler measures, then the target assessment, followed by the CR indices, and lastly the warning manipulation check items. For the conditions with warnings, the order remains the same, except for that there is a warning message displayed at the beginning of the assessment.

Additional Items. I also included several questions that asked participants about their demographic information (age and sex; see appendix xx), college grade-point average, and self-reported standardized test scores. Self-reported standardized test scores and college grade point average are not ideal because participants may either lie about or misremember their score, thereby adding unnecessary measurement error. Despite this methodological concern, self-report college grade point average and self-report standardized test scores still display high convergence with actual grade point average (r = .90) and actual standardized test scores (r = .82; Kuncel et al., 2005). I used self-report measures because obtaining such information from university records would take time that would significantly delay the completion of this work. For any further publication of this work, I plan to obtain and use standardized test score and grade point average data from university records in my analyses.

III. RESULTS

Preliminary Analyses. In the following sections, I describe the preliminary analyses for response time effort, infrequency items, instructed-response items, and long-string analysis. Specifically, I describe and, when necessary, justify the decision process I used when creating each index. Following this section, I then report the analyses and results of each substantive hypothesis.

Response Time Effort. I first investigated whether cutoffs could be set through the visual inspection with conditional response accuracy method or the change in information method. Both methods have traditionally been used in datasets with large sample sizes (e.g., in Wise, 2019, n = 23,000). To explore whether cutoffs could be set with a smaller sample size, I attempted to use both methods to set cutoffs on a random sample of 10 items for each of the 3 tests. For the visual inspection with conditional response accuracy method, the data were too sparse at earlier response times to determine thresholds with confidence across the 10 items in the MAT, SILS-V, and SILS-A. For the change in information method, I was able to set thresholds on the majority of MAT (64 out of 80 items), SILS-V items (28 out of 40), and SILS-A items (12 out of 20).

To investigate the validity of these thresholds, I tested whether the sum of these thresholds correlated with test performance and with thresholds set by the normative threshold method, which sets a threshold at the item-level at a certain percentage (e.g., 20%) of the mean response time. Recent meta-analytic evidence has suggested that the

normative threshold (NT) method has comparable validity to the visual inspection with conditional response accuracy method (Rios & Deng, 2021), and that the NT method can work in smaller samples (Rios et al., 2014). However, this method is not ideal because research has suggested that the typical percentages used (e.g., 10%, 20%) with the NT method were too conservative (Rios & Deng, 2021; Wise, 2019).

To counterbalance this potential weakness, for the MAT and SILS-V, I selected the threshold percentage that identified on average groups responses close to chance. The same criterion has been used in other studies to investigate the validity of thresholds (Soland et al., 2021). I tested three percentages (i.e., 10%, 20%, and 30%) that I have seen used in published work (Soland et al., 2021; Wise, 2019; Wise & Ma, 2012) and the change in information threshold. The assumption is that careless responses are random and should display accuracy rates close to chance (i.e., 25%). A 30% threshold on the MAT and SILS-V identified responses that were the closest to chance (25.4% for MAT; 31.3% for SILS-V) than the other NT and change in information thresholds. For the SILS-A, given there are no response options, I selected the threshold method (tested methods were NT10, NT20, NT30, and change in information method) that best converged with RTE on the MAT and SILS-V, which was NT30 (MAT RTE r = .58; SILS-V RTE r = .53).

Infrequency items. Prior to running any substantive analyses, I scored and investigated the efficacy of each infrequency item within each test. Due to a clerical

error, infrequency items were not included in the SILS-A, so only data for the 10 MAT items and 7 SILS-V items were available. Without dropping items, the 10 MAT infrequency items ($\alpha = .77$) adequate internal consistency, but the 7 SILS-V infrequency items displayed poor internal consistency ($\alpha = .61$). I dropped item 10 for the MAT and items 1 and 3 for the SILS-V from substantive analyses because each item attenuated the internal consistency of the item set. After dropping items, the internal consistency of the MAT and SILS-V infrequency items improved to .78 and .72 respectively.

To further check the validity of retained infrequency items, I used the Rasch model to estimate difficulty parameters for the SILS-V and MAT infrequency items. I estimated parameters separately for (a) the MAT items and MAT infrequency items and (b) the SILS-V items and SILS-V infrequency items. If the infrequency items are working as intended (i.e., functioning as easy test items of ability), then the items should have low difficulty parameters. Results supported this prediction. Both the MAT infrequency items (mean difficulty parameter = -3.50; sd = .52) and SILS-V infrequency items (mean difficulty parameter = -4.46; sd = .35) displayed low difficulty parameters.

Instructed Response Items. I first investigated the efficacy of individual MAT and SILS-V instructed-response items through the same process used for infrequency items. Due to a clerical error, instructed-response items were not included in the SILS-A, so only data for the 4 MAT items and 3 SILS-V items were available. The 4 MAT instructed-response items displayed internal consistency ($\alpha = .84$) without any

modifications. Conversely, the items for SILS-V displayed poor internal consistency (α = .45), and dropping item 3 increased the internal consistency to α = .67.

Long-string analysis. Within each block of the 7 blocks of MAT and SILS-V items, I computed a long-string maximum and a long-string average score. I then created a long-string maximum total score and long-string average total score through summing standardized scores of each block. Given that the long-string maximum and average scores displayed high convergence (r = .86) and to minimize the amount of tests, I conducted an exploratory factor analysis upon the two long-string indices to discern whether the two indices (a) measured the same latent construct and (b) could therefore be combined into a composite. Parallel analysis strongly suggested that the data was composed of 1 factor. Given this evidence of unidimensionality, I standardized and combined the long-string maximum and long-string average total scores into a long-string sum score.

Psychometric Synonyms. To compute psychometric synonyms, I first investigated whether there were sufficiently correlated ($r \geq .6$) item pairs. 1 pair in the dataset was correlated at or higher than .6 (highest r = .61). To compute the index, I took the 20 most highly correlated item pairs (mean r = .41; sd r = .07) . Huang et al., (2012) used a similar method in computing the index.

Hypothesis Testing

In the sections below, to provide context for interpreting the results, I first tested how response time effort (RTE) and the student opinion scale performed under my nomological network. I then described how I tested each hypothesis and summarized the results of each test. I then interpret the results in the discussion section.

Nomological Network: Response Time Effort and Student Opinion Scale. For RTE and the student opinion scale, I tested whether the two were correlated, whether each index displayed moderated the relationships between self-report ACT and test performance, whether each index predicted test performance, whether each index displayed discriminant validity with self-report ACT, and whether each index displayed the warning and scale positioning effects. I intended each of these tests to test the construct validity of a test CR index. The results are summarized in Table 5. Below I described the results of these tests and detailed the steps I took when necessary.

Convergence between Response Time Effort and Student Opinion Scale. RTE and the student opinion scale did not converge well (see table 4a). The student opinion scale was weakly related to RTE on the MAT ($r = .18$; $p < .05$). However, the student opinion scale was not related to RTE on the SILS-V ($r = .06$; $p > .05$) or SILS-A ($r = .06$; $p > .05$). These results immediately suggest 2 possibilities: (1) that one measure functioned poorly or (2) both measures functioned poorly.

Moderation of the Relationship Between Self-Reported ACT Scores and Test Performance. Using moderated regression, I tested whether RTE on the MAT and SILS-

V and the student opinion scale moderated the relationship between self-reported ACT scores and test performance such that the associations were weaker among those who scored as having responded more carelessly on RTE and the student opinion scale. RTE on the MAT (ΔR^2=1.9%; $p < .01$; see figure 1) and SILS-V (ΔR^2=2.1%; $p < .001$; see figure 2) did moderate the associations such that the convergence between test performance and self-reported ACT was lower in those who scored more careless on RTE. Conversely, the student opinion scale neither moderated the association between MAT performance and self-reported ACT scores ($p = .86$) nor the association between SILS-V performance and self-reported ACT scores ($p = .90$).

Moderation of the Relationship Between Self-Reported GPA and Test Performance. I tested whether RTE or the student opinion scale moderated the relationship between self-reported GPA and test performance. Results indicated that RTE on the MAT moderated the association between MAT performance and self-reported GPA (ΔR^2=1.5%; $p < .05$; see figure 3). RTE on the SILS-V also moderated the relationship between SILS-V performance and self-reported GPA (ΔR^2=2.1%; $p < .01$). Both interactions displayed a weaker association at higher levels of CR on RTE and a stronger association on lower levels of CR on RTE. Conversely, the student opinion scale neither moderated the relationship between MAT performance and self-reported GPA ($p = .15$) nor the relationship between SILS-V performance and self-reported GPA ($p = .29$).

Prediction of Test Performance. RTE moderately predicted test performance, whereas the student opinion scale was weakly correlated with test performance. MAT, SILS-V, and SILS-A RTE moderately predicted performance on the MAT (r = -.41; p < .001), SILS-V (r = -.31; p < .001), and SILS-A (r = -.46; p < .001) respectively. The student opinion scale was weakly correlated with performance on the MAT (r = .22; p <. 001), SILS-V (r = .14; p <. 05), and SILS-A (r = .15; p <. 05).

Discriminant Validity with Self-Reported ACT Scores. Both RTE and the student opinion scale displayed discriminant validity with self-reported ACT scores. Self-reported ACT was not significantly related to RTE on the SILS-V (r = .09; p > .05), but was weakly and significantly related to RTE on the MAT (r = .13; p < .05) and SILS-A (r = .14; p < .05). Given the small strength of associations between RTE and self-reported ACT, RTE displayed discriminant validity. The student opinion was not significantly related to self-reported ACT scores (r = .03; p > .05), thereby demonstrating discriminant validity.

Warning Effect. Prior to testing whether those in the warning condition displayed significantly lower rates of CR on RTE and the student opinion scale, I tested for heteroskedasticity. Given that CR indicators are often non-normally distributed, I used the Fligner-Killeen test, which is robust to departures from normality (Algina et al., 1989). To be cautious, I used an alpha level of .20 when testing this assumption. The Fligner-Killeen test suggested that the assumptions of homogeneity of variance was not

violated for the student opinion scale (p =.91), but was violated for SILS-A RTE (p = .13), SILS-V RTE (p = .17), and MAT RTE (p = .03). For the RTE dependent variables, I used the Brown-Forsythe test, which corrects for heteroskedasticity.

For the student opinion scale, using a one-way ANOVA, the warning condition self-reported significantly less carelessness than the control condition (p <.001; d = .49). The warning manipulation did have a significant effect on RTE on the MAT (p = .03; d = .26), where participants in the warning condition displayed significantly less CR than those in the control condition. However, there were no significant differences between the warning condition and the control condition on the SILS-V (p = .65) and SILS-A (p = .14).

Scale Positioning Effect. A significant scale positioning effect was not detected in either RTE on the SILS-V or on the SILS-A. The Fligner-Killeen test indicated that there was heteroskedasticity in the SILS-V (p = .17) and SILS-A (p = .13). Therefore, I used the Brown-Forsythe F test for a scale positioning effect. The results indicated that there were no significant differences between the scale positioning conditions in RTE on the SILS-A (p = .32) or on the SILS-V (p = .81).

Now that I've described analyses for RTE and the student opinion scale (see Table 5 for summary), below I describe the analyses for hypotheses 1-6. Specifically, I restate each hypothesis, describe the process in which I used for each test, and judge

whether each hypothesis was supported. Following this section, I interpret these results in the discussion section.

Hypothesis 1a. For hypothesis 1a, I predicted that infrequency items would converge with RTE and the student opinion scale. MAT infrequency items were strongly and significantly related to MAT RTE (r = .81; see Table 4a) and SILS-V RTE (r = .51). SILS-V infrequency items were significantly related to response time effort on the SILS-V (r = .78) and MAT (r = .47). The student opinion scale was significantly related to MAT infrequency (r = .16; p < .05), but not for SILS-V infrequency items (r = .09; p > .05). Thus, hypothesis 1a was partially supported.

Hypothesis 1b. For hypothesis 1b, I predicted that infrequency items would moderate the relationship between test performance and self-report ACT scores. Using moderated multiple regression, I found that MAT infrequency items did moderate the relationship between MAT performance and self-reported ACT scores (ΔR^2=1.1%, p < .05; see Tables 6a), such that the relationship was lower in participants who missed more MAT infrequency items. However, SILS-V infrequency items did not moderate the relationship between SILS-V performance and self-reported ACT scores (p = .75). Hypothesis 1b was partially supported.

Hypothesis 1c. For hypothesis 1b, I predicted that infrequency itesm on the MAT and SILS-V would moderate the relationship between test performance and self-reported GPA. Neither did the MAT infrequency items significantly moderate the relationship (p =

.45; see Tables 7a), nor did the SILS-V infrequency items moderate the relationship (p = .22). Hypothesis 1c was not supported.

Hypotheses 1d and 1e. Hypothesis 1d and 1e were both supported. MAT infrequency items significantly predicted test performance on the MAT items (r = .41; p < .001). SILS-V infrequency items significantly predicted SILS-V test performance (r = .32; p < .001). Furthermore, MAT and SILS-V infrequency items were neither significantly nor strongly correlated with self-reported standardized test scores (r =.12; p = .053 for MAT; r = .04; p > .5 for SILS-V), thereby supporting hypothesis 1e.

Hypothesis 1f and 1g. Prior to testing hypotheses 1f and 1g, I used the Fligner-Killeen test to determine whether the assumption of homogeneity of variance in analysis of variance (ANOVA) was met using an α level of .20. Results from the Fligner-Killeen test suggested that the assumption homogeneity of variances was not violated for either the test for the warning effect (p = .49) or the scale positioning effect (p = .38). With the assumption upheld, I used a one-way ANOVA to test whether there were significant differences on SILS-V infrequency items between (1) warning conditions and (2) the two scale positioning conditions. The results did not support either hypothesis. Neither did participants in the warning condition significantly get more SILS-V infrequency items right (p =.88; see Table 8a) nor did participants in the condition that completed the SILS first significantly get more SILS-V infrequency items right (p =.64; see Table 9a).

Hypothesis 2a. For hypothesis 2a, instructed-response items on the were not related to RTE on the MAT (r = .09; p = .34; see table 4a), but there was statistically significant, but there was a weak correlation between SILS-V instructed response items and SILS-V RTE (r = .19; p < .05). The student opinion scale was not significantly related to the instructed-response items on the MAT (r = .04; p > .05) and SILS-V (r = .00; p >.05). Given there was only one significant weak association of the four tests, I did not find support for hypothesis 2a.

Hypothesis 2b. For hypothesis 2b, I predicted that instructed-response items on the MAT and SILS-V would moderate the relationship between test performance and self-reported ACT scores such that the associations would be weaker in participants who miss more instructed-response items. Neither MAT instructed-response items nor SILS-V moderated either relationship between self-reported ACT scores and MAT performance (p = .11; see tables 6b) and self-reported ACT scores and SILS-V performance (p = 52). Hypothesis 2b was not supported.

Hypothesis 2c. Hypothesis 2c predicted that instructed response items would moderate the relationship between self-reported GPA and test performance such that the associations would be lower in those who missed more instructed-response items. Neither instructed-response items on the MAT (p = .30; see tables 7b) nor instructed-response items on the SILS-V (p = .27) displayed significant moderation effects. Hypothesis 2c was not supported.

Hypothesis 2d. For hypothesis 2d, I investigated whether instructed-response items were positively correlated with performance on the MAT and SILS-V. Instructed-response items did not predict performance on the MAT (r = .02; p = .72), but there was a significant weak correlation between instructed-response items on the SILS-V and SILS-V performance (r = .19; p <.05). Instructed-response items did significantly predict test performance on the SILS-V, but not the MAT. Therefore, hypothesis 2d was partially supported. .

Hypothesis 2e. For hypothesis 2e, I tested whether instructed-response items displayed discriminant validity with self-reported ACT scores. Instructed-response items were nonsignificantly and weakly related with self-reported ACT scores (r = .01; p > .05) on the MAT and SILS-V (r = .07; p > .05). Hypothesis 2e was supported.

Hypothesis 2f and 2g. Before testing hypotheses 2f and 2g, I used the Fligner-Killeen test to determine whether the assumption of homogeneity of variance in analysis of variance (ANOVA) was met. Results from the Fligner-Killeen test suggested that the assumption homogeneity of variances was not violated for either the test for the warning effect (p = .75) or the scale positioning effect (p = .76). With the assumption upheld, I used a one-way ANOVA to test whether there were significant differences on SILS-V infrequency items between (1) warning conditions and (2) the two scale positioning conditions. The results did not support either hypothesis. Neither did participants in the warning condition significantly get more SILS-V instructed-response items right (p =.59;

see tables 8b) nor did participants in the condition that completed the SILS first significantly get more SILS-V instructed-response items right (p =.44).

Hypothesis 3a. For hypothesis 3a, I predicted that the psychometric synonyms index would significantly and positively converge with response time effort and the student opinion scale. The psychometric synonyms index was not practically related to response time effort on the MAT (r = .02; p = .76; see table 4a) or SILS-V (r = .09, p > .05). Hypothesis 3a was not supported.

Hypothesis 3b. I predicted that psychometric synonyms would moderate the associations between (1) MAT performance and self-reported ACT scores and (2) SILS-V performance and self-reported ACT scores. Neither hypothesis was supported. Neither did psychometric synonyms moderate the relationship between MAT performance and self-report ACT scores (p > .05; see tables 6c) nor did psychometric synonyms moderate the relationship between SILS-V performance and self-reported ACT scores (p > .05).

Hypothesis 3c. I predicted that psychometric synonyms would moderate the associations between (1) MAT performance and self-reported GPA and (2) SILS-V performance and self-reported GPA. Neither did psychometric synonyms moderate the relationship between MAT performance and self-reported GPA (p = .37; see tables 7c) nor did psychometric synonyms moderate the relationship between SILS-V performance and self-reported GPA (p = .75). Hypothesis 3c was not supported.

Hypothesis 3d. For hypothesis 3d, I predicted that psychometric synonyms would predict test test performance on the MAT, SILS-V, and SILS-A. Psychometric synonyms was negatively related to performance on the MAT (r = -.11; p = .053; see table 4c), SILS-V (r = -.13; p = .03), and SILS-A performance (r = -.28; p < .001). Given that these relationships are in the opposite direction, hypothesis 3d was not supported.

Hypothesis 3e. I predicted that psychometric synonyms would display discriminant validity with self-reported ACT scores. There was a small significant relationship between synonyms and self-reported ACT scores (r = -.15; p < .05). Given the practical insignificance of this relationship, hypothesis 3e is supported.

Hypotheses 3f and 3g. Before testing hypothesis 3f, I used the Fligner-Killeen test to determine whether the assumption of homogeneity of variance in analysis of variance (ANOVA) was met. Results from the Fligner-Killeen test suggested that the assumption homogeneity of variances was not violated for either the test for the warning effect (p = .92). With the assumption upheld, I first used a one-way ANOVA to test whether there were significant differences on SILS-V infrequency items between warning and control conditions. The results did not support hypothesis 3f. There was not a significant difference between the warning condition and control condition on psychometric synonyms (p = .08).

For hypothesis 3g, I predicted that psychometric synonyms would indicate greater CR on SILS items in the condition that completes the SILS first than the condition that

completes the SILS-V at the end of the assessment. Prior to testing this hypothesis with a one-way ANOVA, I first used the Fligner-Killeen test to check for heteroskedasticity. The Fligner-Killeen test suggested that there was homogeneity of variance (p = .55). When testing the hypothesis, the results did not support hypothesis 3g.. Participants in the two scale positioning conditions did not significantly differ in psychometric synonym scores (p = .22)

Hypothesis 4a. I predicted that the self-report diligence scale would positively correlate with RTE and student opinion scale. Diligence was weakly related to RTE on the MAT (r =.15; p < .05), but was not related to RTE on the SILS-V (r = 0.05; p > .05) or SILS-A (r = .08; p > .05). Diligence was moderately related with the student opinion scale (r = .61; p <.01). Given the lack of practically significant associations with RTE, hypothesis 4a was partially supported.

Hypothesis 4b. I predicted that the diligence scale would moderate (1) the relationship between MAT performance and self-reported ACT and (2) the relationship between SILS-V performance and self-reported ACT. Using moderated regression, I found that diligence neither moderated (1) the association between MAT performance and self-reported ACT scores (p = .80) nor (2) the relationship between SILS-V performance and self-reported ACT scores (p = .76). Hypothesis 4b was not supported.

Hypothesis 4c. I predicted that the diligence scale would moderate (1) the relationship between MAT performance and self-reported GPA and (2) the relationship

between SILS-V performance and self-reported GPA such that the associations are weaker in participants who self-reported having responded more carelessly to the assessments. I found that diligence neither moderated (1) the association between MAT performance and self-reported GPA scores (p = .72) nor (2) the relationship between SILS-V performance and self-reported GPA scores (p = .74). Hypothesis 4c was not supported.

Hypothesis 4d. I predicted that self-reported diligence would be related to performance on the MAT, SILS-V, and SILS-A. Diligence did predict performance on the MAT (r = .15; p < .05), SILS-V (r = .12; p < .05), and SILS-A (r = .13; p < .05). Hypothesis 4d was supported.

Hypothesis 4e. I predicted that diligence would display discriminant validity with self-reported ACT scores. Diligence was not significantly associated with self-report ACT scores (r =.04; p > .05). Hypothesis 4e was supported.

Hypotheses 4f. I predicted that participants in the warning condition would display significantly higher scores on the diligence measure than participants in the control condition. Prior to testing hypothesis 4f, using the Fligner-Killeen test, I tested whether the variances of condition were not significantly different. Results from the Fligner-Killen test suggested that the assumption of homogeneity of variance was upheld (p = .94). When testing hypothesis 4f, I found that participants in the warning condition

displayed significantly greater diligence scores than the control condition (d = .42; p < .001). Hypothesis 4f was supported.

Hypothesis 5a. Hypothesis 5a predicted that the long-string composite would significantly converge with RTE and the student opinion scale. The long-string composite significantly converged with RTE on the MAT (r = .33; p < .01), SILS-V (r = .42; p < .01), and SILS-A (r =.29; p < .01), but long-string was not significantly related to the student opinion scale (r = -.06; p > .05). Hypothesis 5a is partially supported. .

Hypothesis 5b. I predicted that long-string analysis would moderate the relationships between (1) MAT performance and self-reported ACT scores and (2) SILS-V performance and self-reported ACT scores such that each association would be weaker in those participants who were more careless on long-string. Long-string significantly moderated the relationships between (1) MAT performance and self-reported ACT scores (ΔR^2=3.1%; p <.01). However, the effect was not in the predicted direction. Specifically, the relationship was weaker when participants displayed less carelessness on long-string. Furthermore, long-string did not significantly moderate the relationship between (2) SILS-V performance and self-reported ACT scores (p = .053). Hypothesis 5b was not supported.

Hypothesis 5c. I predicted that long-string would moderate (1) the relationship between MAT performance and self-reported GPA and (2) the relationship between SILS-V performance and self-reported GPA such that the associations are weaker in

participants who score as more careless on long-string. The results did not support hypothesis 4c. Long-string did not significantly moderate the relationship between self-reported GPA and MAT performance (p >.05), but did significantly moderate the relationship between self-reported GPA and SILS-V performance (p <.05; ΔR^2=4.4%). However, the effect was not in the predicted direction. Specifically, the relationship between self-reported GPA and SILS-V performance was weaker among those participants who had displayed less careless responding on long-string. Therefore, hypothesis 5c was unsupported.

Hypothesis 5d. Hypothesis 5d predicted that the long-string composite would significantly predict test performance. Long-string significantly predicted performance on the SILS-A (r = .14; p < .05), and SILS-V (r = .14; p < .05), but not on the MAT (r = .09; p = .14). Hypothesis 5d was partially supported.

Hypothesis 5e. I predicted that long-string analysis would display discriminant validity with self-report ACT scores. Hypothesis 5e was supported. Long-string analysis was not significantly related to self-report ACT scores (r = .04; p = .37).

Hypothesis 5f. I predicted that participants randomly assigned to a condition with a stern warning would display less CR on long-string. I first checked whether assumption of homogeneity of variance was violated using the Fligner-Killeen test. Using an alpha of .20, results indicated that the assumption was violated (p =.15). Therefore, I used the Brown-Forsythe F test to test hypothesis 5f. Results indicated that the warning condition

did not significantly differ from the control condition in long-string (p = .51). Hypothesis 5f was not supported.

Hypothesis 5g. I predicted that participants randomly assigned to a condition where they took the SILS first would display less CR on long-string. I first checked whether assumption of homogeneity of variance was violated using the Fligner-Killeen test. Results from the Fligner-Killeen test suggested that the assumption was not violated (p = .40). I then used a one-way ANOVA to test hypothesis 5g. The results did not support hypothesis 5g. The two scale positioning conditions did not significantly differ in long-string analysis (p = .38).

Hypothesis 6. I predicted that a three-factor structure with careless responding indices will demonstrate the best fit when compared to other models, with infrequency items, instructed-response items, and response time effort loading onto the first factor, self-report diligence and the student opinion scale loading onto a 2nd factor, and the long-string analysis loading onto a third factor. Prior to testing this structure, I dropped instructed-response items and psychometric synonyms from this analysis because each performed poorly within the nomological network. I retained self-report diligence despite the poor performance because self-report diligence performed similarly to the other self-report CR index, namely the student opinion scale, within the nomological network.

Having dropped these two indices, I tested a model with (1) MAT infrequency items, SILS-V infrequency items, MAT RTE, SILS-V RTE, and SILS-A RTE on factor

1, (2) the student opinion scale and diligence on factor 2, and (3) the long-string composite on factor 3 against a set of different models (I refer to this model from hereon as model 1; see Table xx). Model 2 specified that RTE, infrequency items, long-string composite, diligence, and the student opinion scale would load onto 1 common factor. Model 3 specified that (1) infrequency items, RTE, and long-string composite would load onto 1 factor and (2) that diligence and the student opinion scale would load onto a second factor.

Prior to examining the fit for Model 1, given that infrequency and RTE correlated strongly across tests, I specified (1) the MAT and SILS-V RTE measures to be correlated with each other and (2) specified the infrequency and the RTE measures to be correlated with each other. I included these specifications to the other models as well. Model 1 displayed good model fit ($\chi 2$ (11) = 16.79, $p < .01$; CFI = .996, TLI = .991, SRMR = .023, RMSEA = .043). Conversely, Model 2 displayed relatively poorer fit ($\chi 2$ (13) = 140.94, $p < .01$; CFI = .920, TLI = .829, SRMR = .100, RMSEA = .184). Model 3 ($\chi 2$ (12) = 28.85, $p < .01$; CFI = .990, TLI = .976, SRMR = .044, RMSEA = .069) displayed good fit.

I compared the nested models using a $\chi 2$ difference test. Given that Model 1 is more complicated than Models 2 and 3, Model 1 should significantly display better fit. Model 1 significantly displayed better fit than Model 2 ($\Delta\chi 2$ (−2) = 124.15; $p < .01$) and

Model 3 ($\Delta\chi 2\ (-1) = 12.059$; $p < .01$). Given that the revised hypothesized Model 1 displayed the best fit, hypothesis 6 was supported.

VI. DISCUSSION

The purpose of this study was to investigate the validity of novel measures of CR to tests, namely infrequency, instructed-response, psychometric synonyms, long-string, and diligence, using a nomological network. Existing measures (response time effort (RTE) and the student opinion scale) lack either validity (student opinion scale) or flexibility and ease of use (RTE). To elaborate, the student opinion scale has demonstrated poorer validity relative to RTE in empirical tests (e.g., Wise & Kong, 2005). Furthermore, using the self-report method to capture CR potentially introduces sources of bias, such as self-deception, impression management, or faking. Furthermore, using the self-report method requires the tenuous assumption that respondents who responded carelessly to previous items will not respond carelessly to the student opinion scale items. Whereas RTE has demonstrated superior validity to the student opinion scale, the index requires item-level response time information, and computing the index is further complicated by the task of differentiating between the large amount of complex methods (many of which may require large sample sizes; see Wise, 2019 for review) for setting response time cutoffs at the item-level. To find a measure that is simultaneously valid, practical and simple, using a nomological network, I tested 5 novel indices, namely infrequency items, instructed-response items, psychometric synonyms, long-string analysis, and the diligence scale.

The network predicted that valid measures would (1) converge with existing test CR measures (i.e., RTE and the student opinion scale), (2 and 3) moderate relationships between test performance and external criteria (i.e., self-reported ACT and self-reported GPA), (4) predict test performance, (5) display discriminant validity with self-reported ACT scores, (6) show sensitivity to a warning effect, and (7) show sensitivity to a scale positioning manipulation. Below, I address some discrepancies between the expected and observed performance of nomological network components, specifically in (1) convergence with the student opinion scale, (2) sensitivity to warning effect, and (3) sensitivity to a scale positioning effect. Following this discussion, I then interpret how each index performed under the nomological network, starting with infrequency items and ending with the diligence scale.

Convergence with Response Time Effort and Student Opinion Scale. Prior to discussing the convergent validity of the novel indices, I must address why RTE did not display adequate convergence with the student opinion scale. This result leaves 3 possibilities, that RTE did not function well as a measure of CR, that the student opinion scale did not function well, or that neither index functioned well.

Given the pattern of results, it's clear that the low convergence is due to the poor performance of the student opinion scale. RTE (1) strongly converged with infrequency items, (2) moderately converged with long-string, (3) moderately predicted test performance across tests, and (4) displayed discriminant validity with self-reported ACT

scores. These results replicate and extend previous research supporting the validity of RTE as a measure of CR (see Wise, 2015). Given this support, I interpret convergence with RTE as supporting evidence for validity of novel indices. In contrast, the student opinion scale (1) did not converge well with other CR indices (i.e., RTE, infrequency, long-string), (2) weakly predicted test performance, (3) did not moderate relationships between test performance and external variables, and (4) demonstrated moderate sensitivity to the warning manipulation. The last piece of evidence seems enticing, but the results from the warning manipulation are misleading (see below). Given the weak performance across network components, I do not interpret low convergence with the student opinion scale to indicate poor validity of the novel indices. Therefore, I limit my discussion of convergence with existing test CR indices to convergence with RTE.

Sensitivity to Warning Effect. Only three measures displayed sensitivity to a warning effect in the predicted direction, namely RTE on the MAT (but not on the SILS-V/SILS-A) and the two self-report measures of CR (i.e., diligence and student opinion scale). Furthermore, the warning did not have an effect on test performance across the three tests. These results, combined with the pattern of small, nonsignificant results across the non-self-report CR indices, indicates that the warning effects likely induced response distortion among participants on the self-report CR measures, but did little to induce attentive responding. This possibility, combined with the fact that the manipulation check items for the warning manipulation displayed only a moderate

difference between the control and warning group (Hedge's g = .38), leads me to interpret insensitivity to the warning effect to not reflect poorly on validity of a CR measure. The poor performance of the warning manipulation in this data collection contradicts much previous research in surveys testing warning manipulations (e.g., Gibson & Bowling, 2019; Huang et al., 2012; Bowling et al., 2021).

Sensitivity to Scale Positioning Manipulation. I predicted that each measure would display sensitivity to a scale positioning manipulation such that participants who take the SILS-V after the MAT will display greater CR than those who take the SILS-V before the MAT. None of the study measures displayed a significant scale positioning effect. These results have suggested two possibilities: (1) that participants successfully maintained effort throughout the lengthy assessment or (2) that the manipulation was ineffective and the analysis was therefore not sensitive to detecting diminishing effort among participants.

Assuming the first possibility, this result contradicts previous findings in survey (e.g., Bowling et al., 2020) and test (e.g., Weirich et al., 2017) CR research, but echoes findings in test-fatigue research (Ackerman & Kanfer, 2009). Furthermore, this interpretation implies that researchers and practitioners may be able to use lengthy tests in low-stakes settings without fear of increasingly inducing CR as participants progress through the lengthy test.

Concerning the second possibility, something to note is that the scale positioning manipulation has not been tested in tests and that past research focused on examining item positioning effects on item response theory model parameter estimates (e.g., Weirich et al., 2017; Zeller et al., 2017). Given these analyses require large sample sizes for proper statistical power, I did not replicate their analyses within this study and instead chose a manipulation more suitable for the sample size I was able to collect.

Another possibility is that participants may have found these assessments fun and challenging, which would explain why participants did not get increasingly careless as they progressed throughout the assessment. At the end of the assessment after debriefing, I included an open response question for participants to provide feedback on the study. Several participants noted that they "enjoyed" the study or that it was a "great experience" (see Table 12).

Performance of Novel Measures Under the Nomological Network.

I predicted that each novel index should (1) converge with existing CR indices, (2) predict test performance, (3) moderate the convergent and predictive validity of tests, (4) display discriminant validity with high-stakes ability scores, (5) show sensitivity to a warning effect and (6) scale positioning effect. After examining the performance of certain nomological network components, I concluded that the student opinion scale demonstrated weak construct validity, and that neither the warning nor the scale positioning manipulation induced an effect on careless responding. Therefore, results on

each of these nomological network components are uninformative concerning the construct validity of a measure. Thus, I decided to just judge the validity of each novel index by examining whether the index (1) displayed convergence with RTE, (2) predicted test performance, (3) moderated the convergent and predictive validity of tests, and (4) displayed discriminant validity.

I retained the first two elements because (1) RTE displayed strong validity across nomological network elements, (2) RTE and other CR indices did predict test performance, and test scores displayed good validity (i.e., test scores converged with each other and self-reported ACT scores), thereby indicating that test performance was functioning as a valid criterion. I retained the third element because detection of an interaction effect of CR despite the low prevalence of CR should reflect well on the validity of a CR index. Lastly, I retained discriminant validity with self-reported ACT scores because self-reported ACT scores replicated results demonstrated from previous research (i.e., high convergence with test performance across tests and weak convergence with RTE and SOS), thereby indicating validity for the index. Below, using this amended nomological network, I interpret the results concerning each of the tested indices. I then discuss the implications that each pattern of results has on further research and practice.

Infrequency. Overall, infrequency items performed well under the amended nomological network. Infrequency (1) displayed strong convergence with RTE, (2) moderately predicted test performance, (3) displayed discriminant validity with self-

reported high stakes test scores, but did only moderated 1 of 4 relationships between test performance and external variables. These findings suggest that researchers may use easy test items as measures of CR. Given that non-adaptive tests are likely to administer such items to participants, it's possible that researchers and practitioners could repurpose existing east test items as CR measures after collecting data without prior planning, thereby bolstering the applicability of the infrequency approach. Furthermore, in adaptive tests, such items could be administered to assess whether a participant is responding carelessly.

Instructed-Response. Across the nomological network components, instructed-response performed extremely poorly. Instructed-response items displayed no convergence with RTE, did not moderate relationships between test performance and external variables, and only weakly predicted test performance on the SILS. This poor performance could be due to the (1) confusing nature of instructed-response items and (2) the distinctive appearance of instructed-response items. Concerning the former, participants may be confused by such items and try to solve the item as a normal ability test question (i.e., thereby inflating the Type I error rate), which may have happened in the MAT instructed-response questions. The MAT instructed-response items displayed high flag rates, with just under half of participants missing two or more items of the four items. When designing the items, I had sacrificed clarity of instructions to ensure that the appearance of the items was not different from other MAT items because I was

concerned that careless participants would note the distinctive appearance would stand out to careless participants, who would stop and examine such items (i.e., thereby inflating the Type II error rate).

Given that the SILS-V instructed-response items had clearer instructions and performed better across the nomological network, it's likely that the instructions of the MAT instructed-response items confused participants. In comparison, the three SILS-V items displayed lower flag rates, and displayed stronger convergence with test performance and RTE than MAT instructed-response items. While these results are relatively promising compared to other indices, when taken together, I see no reason to use instructed-response items over infrequency items. Unlike instructed-response items, infrequency items (1) displayed greater validity, (2) potentially assess CR without the knowledge of the participant, and (3) are equally practical to instructed-response items.

Psychometric Synonyms. Psychometric synonyms displayed no promise as a measure of CR to tests. First, I found no item pairs within the 140-item data set that exhibited a correlation above the conventional cutoff (i.e., $r = .6$), which researchers have successfully used with survey data. This dearth of item pairs may be due to differences in scoring between survey and test items. A typical survey item has much greater variability (e.g., range of 1-7 on a 7-point Likert scale) than the dichotomously scored test items in this study (i.e., scored as 0 or 1), which carry the least amount of information a covariate can have. As research has demonstrated, restriction of range has an attenuating effect on

correlations (MacCallum et al., 2002; Schmitt, et al., 2007). Taken together, the failure to identify correlated item pairs may be due to restriction of range within the correlates.

Even when I relaxed the cutoff to $r = .35$ to include the 20 most highly correlated item pairs, psychometric synonyms performed extremely poorly. Psychometric synonyms (1) did not converge with infrequency items or RTE, (2) did not moderate relationships between test performance and external criteria, and (3) only moderately predicted performance on the SILS-A. Taken together, the results strongly suggest that researchers cannot extend psychometric synonyms to capture CR to tests.

Diligence. Diligence displayed unacceptable validity as a CR index, which was a similar pattern of results to that of the student opinion scale. Diligence (1) did not correlate well with RTE, (2) did not moderate the relationships between test performance and external criteria, and (3) weakly predicted test performance. These poor results may indicate that self-report measures provide information about participants at different levels of careless responding. Most CR measures (e.g., RTE, infrequency) are explicitly designed to differentiate a critical point in the continuum that separates those participants who are responding completely without regard to item content and those participants who are expending the minimal amount of effort needed to respond to an item. Conversely, self-report measures may reflect individual differences in effort across continuum instead at the critical juncture. This reasoning could explain the poor convergence with other

measures of CR (i.e., RTE, infrequency, long-string), but can not explain the poor performance on other criteria.

A more plausible and parsimonious explanation is that the poor results of the self-report measures may indicate a problem endemic to the self-report method when applied to measure CR, which is contaminated by response distortion (i.e., impression management and self-deception) and requires the tenuous assumption that careless participants would suddenly stop and attentively respond to a self-report measure. Regardless of the explanation, the results of diligence and the student opinion scale serve as a condemnation of self-report CR indices. Researchers and practitioners should use other alternatives over these measures.

Long-string. Long-string displayed poor validity as a measure of CR to tests. Long-string only moderately converged with RTE across the three tests, did not significantly moderate any relationships between test performance and external criteria, and (2) did not predict test performance. The moderate convergence with RTE and infrequency could be due to the fact that long-string only captures straight-lining, which is just one of several careless response patterns. However, such an explanation cannot explain the poor performance on the 2nd and 3rd components. Long-string did display discriminant validity with self-reported high stakes test scores, which indicates that the measure did not capture individual differences in ability.

The way that items were presented may have negatively impacted the viability of long-string. In survey research, where long-string measures are used to capture CR (e.g., Bowling et al., 2020), researchers look at long-string scores of each participant on a page of items. In such an item presentation, participants might be more likely to engage in straight-lining. Conversely, in my study, I presented items individually to participants, which may discourage straight-lining and instead encourage other careless response patterns. Thus, it's possible that the performance of long-string in the current study was negatively impacted by the methods, and that the long-string analysis may be more valid in assessments where items are displayed on pages instead of individually.

Despite this poor performance relative to infrequency and RTE, what makes long-string potentially useful to researchers and practitioners is the flexibility of the index, which is long-string's greatest strength. The index does not require any prior planning from researchers because researchers can use long-string analyses on any dataset with unscored responses. Furthermore, improved indices of response invariance (e.g., individual response variance; see Dunn et al., 2018) may display ever greater validity while retaining similar or greater flexibility. Thus, despite the relatively poor validity, long-string still may have practical utility due to the flexibility of the analysis. Lastly, the index is useful for any research focused on straight-lining specifically (as opposed to capturing CR broadly; e.g., DeSimone & Harms, 2018).

Taken together, long-string demonstrated poor validity within this study, but researchers may find some use in the measure due to its practicality. Future research should investigate (1) whether long-string works in different kinds of ability assessments (e.g., knowledge tests such as the informational literacy test; see Wise & Kong, 2005) and (2) whether the validity of long-string depends on the presentation method for items (i.e., does validity of the index vary when items are displayed on pages versus displayed individually). To elaborate, long-string may be more effective in tests where multiple items are presented on pages, where participants can respond in strings more seamlessly.

Factor Structure of Careless Responding. A three-factor model that contained (1) infrequency and RTE on one factor, (2) self-report indices on a second factor, and (3) long-string on its own factor fit the data better than a one-factor and two-factor model. These results echo survey research (e.g., Meade & Craig, 2012) that found evidence suggesting that self-report CR indices and measures of straight-lining (i.e., long-string) load onto a separate factor from other CR indices.

Practical Implications. This work has 2 notable practical implications. First, that easy test items and long-string analysis may be applicable to capturing CR to tests. This finding is relevant because the process obtaining the information needed for each index is more flexible than obtaining item-level response time information for RTE, which, as it stands, is the only validated measure of CR to tests. For long-string, researchers just need to know the order in which items were presented. For infrequency, researchers just need

to include extremely easy test items in their data collection. Both indices are simpler than RTE and are usable in tests where item-level response time was not or cannot be collected (e.g., pen-and-paper tests).

Second, this work neither found support for a warning manipulation nor found support for a scale positioning manipulation. The former result suggested warnings may not be an effective preventative technique. The latter result suggested that participants did not get progressively careless throughout the lengthy assessment. Therefore, researchers and practitioners may be able to administer assessments of similar lengths to participants without fear of inducing greater CR.

Limitations. This study has several limitations. First, given the small sample size (n = 291) and the low prevalence of CR within the dataset (which may have reduced the effect of the moderation), it's possible that some tests for moderation were underpowered. Furthermore, for the purposes of recording item-level response time information, I displayed test items to participants individually. This order of presentation may have discouraged straight-lining, and encouraged random responding. Therefore, the results for the long-string index may not be generalizable to other test administration contexts. Furthermore, this study used self-report measures of standardized test scores and grade point average, which are not completely accurate methods for obtaining either source of information. However, self-reported standardized test scores and GPA do converge strongly with actual standardized test scores and GPA (see Kuncel et al., 2006).

Furthermore, self-reported ACT scores converged with test performance on the MAT (r = .56), SILS-V (r = .43), and SILS-A (r = .48), which further supports interpreting the self-report measure as an indicator of ability.

Future Research. I outline 6 possible directions for future research. First, future research should replicate the current findings and extend them to populations that have different sample characteristics from student samples (e.g., crowd sourced samples, grade-school students, job incumbents in concurrent validation designs). Second, future research should investigate the incremental validity of infrequency over RTE using an appropriate criterion (e.g., a criterion similar to item content recognition, which is in survey research; see Bowling et al., 2022). Such a finding would further support the validity of long-string and infrequency. Third, research should investigate whether practitioners can identify and utilize easy test items to successfully measure CR after the data is collected. This possibility would bolster the practical utility of the infrequency approach.

Fourth, researchers should investigate whether the infrequency approach and long-string analysis extends to other ability tests (e.g., nonverbal assessments such as Raven's Progressive Matrices) and knowledge tests (e.g., information literacy test (Cameron et al., 2007), situational judgment test, job knowledge test). Fifth, research should investigate whether long-string analysis displays variability in validity depending

on item-presentation order. Long-string may demonstrate greater promise when items are displayed on pages, where participants may be more likely to engage in straight-lining.

Sixth, researchers should investigate the role of self-efficacy in the relationship in careless responding to tests. Self-efficacy describes an individual's belief regarding their ability to achieve a given goal (Heslin & Klehe, 2006). Participants low in self-efficacy, who doubt their ability to accomplish the task at hand, may be more liable to engage in careless responding as a means of avoiding engagement with the task, which would result in a negative relationship between self-efficacy and task performance. Therefore, careless responding may serve as a mechanism through which self-efficacy affects task performance.

Practical Recommendations. Given the results of my study, I recommend the use of RTE and easy test items as measures of careless responding to tests. Long-string may have some use in detecting extreme cases of straight-lining, but I would caution using the index beyond that niche purpose. Furthermore, given the poor results of the self-report measures and psychometric synonyms, I would not recommend using either self-report measures or psychometrics synonyms to assess careless responding.

Conclusion. The purpose of this work was to investigate the construct validity of 5 measures (i.e., infrequency, instructed-response, psychometric synonyms, diligence, and long-string) of CR to tests through using a nomological network. The results found (1) strong support for infrequency items, (2) inconsistent support for instructed-response

items, (3) no support for psychometric synonyms, (4) little support for the diligence scale, and (5) moderate support for long-string analysis as measures of CR to tests. This study adds to the literature by introducing two new measures (i.e., infrequency and long-string) that compensate for the weaknesses of the only valid indicator of CR to tests, namely response time effort. Given the results of the current study, researchers and practitioners should further explore the efficacy of using infrequency and long-string to measure CR on tests.

V. REFERENCES

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163–181. https://doi.org/10.1037/a0015719

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in PreK–12 education. Educational Researcher, 29, 24–25.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing (6th ed.). Washington DC: American Educational Research Association

Bäckström, M., & Björklund, F. (2019). Is reliability compromised towards the end of long personality inventories? European Journal of Psychological Assessment, 35, 14–23. https://doi.org/10.1027/1015-5759/a000363

Barrett, G. V., Phillips, J. S., Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 66, 1-6.

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

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe,K. (1992).MMPI-2 random responding indices:Validation using a self-report methodology. Psychological Assessment, 4, 340=345.

Berinsky AJ, Margolis MF and Sances MW (2013) Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. American Journal of Political Science 58: 739–753.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Within-person increases in careless responding during questionnaire completion. *Organizational Research Methods*. doi: 10.1177/1094428120947794

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016).

Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*, 218-229.

Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In D. H. Saklofske (Ed.), The SAGE handbook of personality theory

and assessment. Vol. 2: Personality measurement and testing (pp. 179–198). Thousand Oaks, CA: Sage

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data.

Journal of Experimental Social Psychology, 66, 4-19.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. Educational Assessment, 12(1), 23–45. doi:10.1080/10627190709336946

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. Applied Psychology: An International Review, 67, 309–338.

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. Journal of Business and Psychology, 33, 559–577.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36,* 171-181.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. Proceedings of the National

Academy of Sciences of the United States of America, 108, 7716–7720. doi:10.1073/pnas.1018601108

Ehlers, C., Greene-Shortridge, T. M., Weekley, J. A., & Zajack, M. D. (2009). The exploration of statistical methods in detecting random responding. Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology, Atlanta, GA.

Eisenthal, S., & Harford, T. (1971). Correlation between the Raven Progressive Matrices Scale and the Shipley Institute of Living Scale [Abstract]. Journal of Clinical Psychology, 27(2), 213-215.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. Assessment in education. Principles, Policy & Practice, 17, 345–356.

Evans, D., Boggero, I., & Segerstrom, S. (2015). The nature of self-regulatory fatigue and "ego depletion": Lessons from physical fatigue. Personality and Social Psychology Review. http://dx.doi.org/10.1177/1088868315597841. 2015.

Francavilla, N. M., Meade, A. W., & Young, A. L. (2019). Social interaction and internet-based surveys: Examining the effects of virtual and in-person proctors on careless response. Applied Psychology, 68, 223-249.

Frey MC, Detterman DK. Scholastic Assessment or g?: The Relationship Between the Scholastic Assessment Test and General Cognitive Ability. *Psychological Science*. 2004;15(6):373-378. doi:10.1111/j.0956-7976.2004.00687.x

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. Journal of Official Statistics, 22, 313–328.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opinion Quarterly, 73, 349–360.

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*, 410-420. doi: 10.1027/1015-5759/a000526.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23(1), 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. Public Opinion Quarterly, 45, 549–559.

Heslin, P.A., & Klehe, U.C. (2006). Self-efficacy. In S. G. Rogelberg (Ed.), Encyclopedia of Industrial/Organizational Psychology (Vol. 2, pp. 705-708). Thousand Oaks: Sage.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2014). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299-311.

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

Huang, J. L., & DeSimone, J. A. (2020). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology*.

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

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: A meta-analysis. Psychological Bulletin, 136, 495–525. http://dx.doi.org/10.1037/a0019486.

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

Kam, C. C. S., & Chan, G. H. H. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences, 129,* 83-87.

Kim, D.S., McCabe, C.J., Yamasaki, B.L., Louie, K.A., King, K.M., 2017. Detecting random responders with infrequency scales using an error balancing threshold. Behav. Res. Methods. http://dx.doi.org/10.3758/s13428-017-0964-9.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213-236.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All? *Journal of Personality and Social Psychology, 86*(1), 148–161. https://doi.org/10.1037/0022-3514.86.1.148

Liu M, Wronski L. Trap questions in online surveys: Results from three web survey experiments. International Journal of Market Research. 2018;60(1):32-49. doi:10.1177/1470785317744856

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83.

Martin, J. D., Blair, G. E., Stokes, E. H., & Lester, E. H. (1977). A validity and reliability study of the Slosson Intelligence Test and the Shipley Institute of Living Scale. Educational and Psychological Measurement, 37(4), 1107-1110.

McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior, 84*, 295-303.

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

Meade, A. W., & Pappalardo, G. (2013). Predicting careless responses and attrition in survey data with personality. In 28th Annual Meeting of the Society for Industrial and Organizational Psychology, Houston, TX.

Pearson Assessments, Meagher, D. G., Pan, T., Wagner, R., & Miller, J. R. (2021). MAT Reliability and Validity. http://images.pearsonassessments.com/Images/dotCom/milleranalogies/pdfs/MAT_Reliability-Validity_FNL.pdf

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT based common item equating design. Applied Measurement in Education, 22(1), 38-60. https://doi.org/10.1080/08957340802558342

Nichols, A. L. & Edlund, J. E. (2020) Why don't we care more about carelessness? Understanding the causes and consequences of careless participants, International Journal of Social Research Methodology, 23:6, 625-638, DOI: 10.1080/13645579.2020.1719618

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. Journal of Clinical Psychology, 45, 239 –250.

LearningExpress Editors, Dermott, B. D., Gade, S. G., McLean, K. M., Recco, W. R., & Schultz, C. S. (2002). 501 Word Analogy Questions (1st ed.). Learningexpress, LLC.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45, 867–872.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. Frontiers in Psychology, 1, 220:1–7. doi:https://doi.org/10. 3389/fpsyg.2010.00220

Periard, D. A., & Burns, G. N. (2014). The relative importance of Big Five Facets in the prediction of emotional exhaustion. Personality and Individual Differences, 63, 1–5.

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

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. Journal of Educational Measurement, 34, 213–232.

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web survey panels over time. Survey Research Methods, 9(2). https://doi.org/10.18148/srm/2015.v9i2.6128

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2014). *Motivation in education: Theory, research and applications*.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology: Interdisciplinary and Applied, 9,* 371–377. https://doi.org/10.1080/00223980.1940.9917704

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge University Press.

Sundre, D. L. (2007). The student opinion scale: A measure of examinee motivation. Retrieved

from James Madison University, Center for Assessment and Research Studies website: http://www.jmu.edu/assessment

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? Contemporary Educational Psychology, 29, 6–26. doi:10.1016/S0361-476X(02)00063-2

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. Assessment Update, 14(1), 8–9.

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, 129–151. doi:10.1353/jge.0.0047

Maggie E. Toplak, Richard F. West & Keith E. Stanovich (2014) Assessing miserly information processing: An expansion of the Cognitive Reflection Test, Thinking & Reasoning, 20:2, 147-168, DOI: 10.1080/13546783.2013.844729

Vohs, K. D., & Faber, R. (2007). Spent resources: Self-regulatory resource availability affects

impulse buying. Journal of Consumer Research, 33, 537–547 http://dx.doi.org/10.1086/510228.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless

responding during online surveys. *Applied Psychology: An International Review*, *67*, 231-263.

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize

careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554-568.

Weiss, J. L., & Schell, R. E. (1991). Estimating WAIS—R IQ from the Shipley Institute of Living Scale: A replication. *Journal of Clinical Psychology*, *47*(4), 558–562. https://doi.org/10.1002/1097-4679(199107)47:4<558::AID-JCLP2270470414>3.0.CO;2-W

Woods, C. M. (2006). Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 189–194. https://doi.org/10.1007/s10862-005-9004-7

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. Applied Psychological Measurement, 41(2), 115-129. https://doi.org/10.1177/0146621616676791

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, 93–112.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in

low-stakes testing programs. The Journal of General Education, 58(3), 152–166.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. Applied

Measurement in Education, 28, 237–252.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. Educational Measurement: Issues and Practice, 36(4), 52–61. doi:10.1111/emip.12165

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. Applied Measurement in Education, 32(4), 325–336. doi:10.1080/08957347.2019.1660350

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

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effortmoderated

IRT model. Journal of Educational Measurement, 43, 19–38.

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

Wise, S. L., & Ma, L. (2012, April). Setting response time thresholds for a CAT item pool:The normative threshold method. Paper presented at the Annual Conference of the National Council on Measurement in Education, Vancouver, Canada.

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

Zhang, C. & Conrad, F. (2014). Speeding in web surveys: the tendency to answer very fast and its association with straightlining. Survey Research Methods, 8(2), 127–135.

Table 1a

Table Describing Careless Responding Measures

Index	Description	Mode	Advantages	Limitations
Infrequency Items	Items that are designed to have low response variance from attentive respondents so that the item differentiates between careless and attentive respondents.	Survey	Items are easy to make and use. Items display good validity in capturing CR in surveys.	Requires inclusion of items, which necessitates prior planning and adds to participant burden. Research isn't clear on how to obtain classification cutoffs, which inhibits practical implementation. Items differ greatly in flag rate and there is no research investigating these differences.
Instructed-Response Items	Items that direct attentive participants to certain responses so that these items just differentiate between careless and attentive respondents.	Survey	Items are easy to make and use. Format of instructed-response items is easily adaptable to different assessment formats. Items show good validity in capturing survey CR.	Requires inclusion of items, which necessitates prior planning and adds to participant burden. Items display heterogeneity in flag rate and there is no research investigating why. No research on setting classification cutoffs, which inhibits practical implementation.
Response Time Effort	Examines carelessness at the item level through investigating a variety of data (see Wise, 2019).	Test	Thus far, the most valid metric for capturing CR to tests. Captures CR without the knowledge of the participant. Many different empirically validated cutoff methods, which allows for great flexibility.	Requires item-level response time information, often which is practically unobtainable.

Table 1b

Table Describing Careless Responding Measures

Index	Description	Mode	Advantages	Limitations
Page time	Investigates carelessness at the page level with response time cutoffs.	Survey	Good evidence supporting the use of page time to capture CR in surveys. Captures CR without the knowledge of the participant.	No research on classification cutoffs, which inhibits practical implementation. Requires page-level response time information, which may be practically unobtainable.
Student Opinion Scale	Self-report scale that assesses how much effort participants expended in test responding and how important test results are to the participant.	Test	Easy to use and apply--practically flexible. Exhibits some validity (but less than response time effort) in capturing CR to tests. There is research on classification cutoffs.	No research distinguishing scores from sources of bias such as impression management or self-deception. Requires the tenuous assumption that careless participants are going to stop responding carelessly and attentively and truthfully respond to the scale.
Diligence	Self-report scale that assesses how effortful participants were in responding to a survey.	Survey	Easy to use and apply--practically flexible. Exhibits some validity (but less than other indicest) in capturing CR to surveys.	No research distinguishing scores from sources of bias such as impression management or self-deception. Requires the tenuous assumption that careless participants are going to stop responding carelessly and attentively and truthfully respond to the scale. No research on classification cutoffs, which inhibits practical implementation.

Table 1c

Table Describing Careless Responding Measures

Index	Description	Mode	Advantages	Limitations
Long-String	*Post hoc* analysis that examines consecutive strings of responses.	Survey	Captures a specific careless response pattern--straight-lining, which may be useful if an investigation necessitates the isolation of this response-pattern. The analysis is extremely flexible--it can be run on any archival dataset that has multiple response options.	Displays poor convergence with other CR indices and does not load onto the same factors. This is likely because the index only captures straight-lining.
Psychometric Synonyms	Examining correlation within-person between highly correlated pairs of items to deduce whether a participant was careless or attentive.	Survey	Good validity evidence supporting the use of synonyms to capture CR in surveys. The analysis is extremely flexible--it can be run on any archival dataset that has correlated item pairs.	No evidence on the appropriate cutoff for pair inclusion or on classification cutoffs. Both gaps inhibit practical implementation. Using the index is certain analyses (e.g., correlation) treats logarithmic information as linear.

Table 2

Nomological Network Table For Capturing Careless Responding to Tests

External Variables or Effects	Proposed Test Careless Responding Indices
Convergence with Student Opinion Scale Hypotheses 1a, 2a, 3a, 4a, and 5a	+
Convergence with Response Time Effort Hypotheses 1a, 2a, 3a, 4a, and 5a	+
Convergence with Test Performance Hypotheses 1d, 2d, 3d, 4d, and 5d	+
Convergence with Archival Standardized Test Scores Hypotheses 1e, 2e, 3e, 4e, 5e	NS
Filtering Effect Hypotheses 1b-5b and 1c-5c	>
Warning Effect Hypotheses 1f, 2f, 3f, 4f, 5f	<
Scale Positioning Effect Hypotheses 1g, 2g, 3g, 4g, 5g	>

Note. + Positive significant relationship. NS hypothesized nonsignificant relationship. >* suggests that the proposed index should display an increase in a correlation after filtering. <* The condition with the manipulation should display less CR than the control condition.

Table 3

Order of Assessments

Condition	Block 1	Block 2	Block 3	Block 4	Block 5
1		Target Assessment	Filler Test Measures	Careless Indices	Warning Manipulation Check
2		Filler Test Measures	Target Assessment	Careless Indices	Warning Manipulation Check
3	Warning Message	Target Assessment	Filler Test Measures	Careless Indices	Warning Manipulation Check
4	Warning Message	Filler Test Measures	Target Assessment	Careless Indices	Warning Manipulation Check

Note. Presentation of different measures within each block was randomized.

Table 4a

Means, Standard Deviations, and Correlations with Response Time Effort

Variable	*M*	*SD*	4	5	6
4. MAT RTE	75.99	10.63			
5. SILS-V RTE	39.50	2.41	.58** [.50, .65]		
6. SILS-A RTE	19.12	2.48	.58** [.50, .66]	.39** [.28, .48]	
7. MAT INF	8.63	1.07	.81** [.76, .84]	.51** [.42, .59]	.55** [.46, .62]
8. SILS-V INF	4.92	0.44	.47** [.38, .56]	.78** [.73, .82]	.18** [.07, .29]
9. MAT IR	2.15	1.64	.10 [-.02, .21]	.08 [-.04, .19]	.12* [.00, .23]
10. SILS-V IR	2.94	0.26	.20** [.09, .31]	.31** [.21, .41]	.19** [.08, .30]
11. PSYN	0.53	0.28	-.02 [-.13, .10]	.09 [-.03, .20]	-.08 [-.19, .04]
12. Long-String Total	-0.00	3.52	.34** [.23, .43]	.42** [.33, .51]	.29** [.18, .39]
13. IRV	-0.00	4.29	.57** [.49, .65]	.59** [.50, .66]	.49** [.40, .57]
14. Diligence	57.01	11.11	.16** [.04, .27]	.05 [-.06, .17]	.09 [-.03, .20]
15. SOS	37.97	6.33	.14* [.02, .25]	.07 [-.04, .19]	.05 [-.07, .16]

Note. M and *SD* are used to represent mean and standard deviation, respectively. MAT, SILS-V, and SILS-A refer to Miller Analogy Test practice items, Shipley Institute of Living Scale Verbal subscale, and Shipley Institute of Living Scale Abstraction subscale respectively. INF and IR refer to infrequency and instructed-response respectively. SOS refers to student opinion scale. PSYN refers to psychometric synonyms. Values in square brackets indicate the 95% confidence interval for each correlation * indicates $p < .05$. ** indicates $p < .01$.

Table 4b

Correlations between Careless Responding Indicators

Variable	1	2	3	4	5	6	7	8
1. MAT INF								
2. SILSV INF	.46** [.37, .55]							
3. MAT IR	.16** [.04, .27]	.06 [-.05, .17]						
4. SILS-V IR	.21** [.09, .31]	.32** [.21, .42]	.16** [.05, .27]					
5. PSYN	-.00 [-.12, .11]	.10 [-.01, .22]	.00 [-.11, .12]	-.04 [-.15, .08]				
6. LST	.31** [.20, .41]	.29** [.18, .39]	.04 [-.07, .16]	.27** [.16, .38]	.05 [-.06, .17]			
7. IRV	.55** [.47, .63]	.45** [.36, .54]	.04 [-.08, .15]	.42** [.32, .51]	.04 [-.07, .16]	.67** [.60, .73]		
8. Diligence	.21** [.10, .32]	.09 [-.02, .21]	.17** [.05, .28]	-.04 [-.15, .08]	.02 [-.09, .14]	-.08 [-.20, .03]	.01 [-.10, .13]	
9. SOS	.16** [.04, .27]	.09 [-.02, .21]	.04 [-.08, .16]	.00 [-.12, .12]	-.01 [-.13, .11]	-.07 [-.18, .05]	.05 [-.06, .17]	.62** [.54, .68]

Note. M and *SD* are used to represent mean and standard deviation, respectively. MAT, SILS-V, and SILS-A refer to Miller Analogy Test practice items, Shipley Institute of Living Scale Verbal subscale, and Shipley Institute of Living Scale Abstraction subscale respectively. INF and IR refer to infrequency and instructed-response respectively. SOS refers to student opinion scale. PSYN refers to psychometric synonyms. LST refers to long-string total. Values in square brackets indicate the 95% confidence interval for each correlation * indicates $p < .05$. ** indicates $p < .01$.

Table 4c

Correlations with Test Performance

Variable	*M*	*SD*	1	2	3
1. MAT performance	36.62	11.58			
2. SILSV performance	26.70	4.83	.71** [.64, .76]		
3. SILSA performance	9.25	3.17	.60** [.52, .67]	.48** [.39, .57]	
4. MAT RTE	75.99	10.63	.41** [.31, .50]	.37** [.27, .46]	.40** [.30, .49]
5. SILS-V RTE	39.50	2.41	.21** [.09, .31]	.35** [.24, .44]	.19** [.07, .30]
6. SILS-A RTE	19.12	2.48	.31** [.21, .41]	.32** [.21, .42]	.46** [.37, .55]
7. MAT INF	8.63	1.07	.39** [.29, .49]	.36** [.25, .46]	.41** [.31, .50]
8. SILS-V INF	4.92	0.44	.14* [.02, .25]	.28** [.17, .39]	.15* [.03, .26]
9. MAT IR	2.15	1.64	.02 [-.09, .14]	.08 [-.04, .19]	.12* [.00, .23]
10. SILS-V IR	2.94	0.26	.04 [-.07, .16]	.19** [.08, .30]	.19** [.07, .30]
11. PSYN	0.53	0.28	-.11 [-.23, .00]	-.13* [-.24, -.02]	-.28** [-.39, -.17]
12. Long-string Total	-0.00	3.52	.06 [-.06, .17]	.10 [-.01, .22]	.12* [.01, .24]
13. IRV	-0.00	4.29	.32** [.22, .42]	.36** [.26, .46]	.29** [.18, .39]
14. Diligence	57.01	11.11	.15**	.12*	.13*

			[.04, .26]	[.01, .24]	[.02, .24]
15. SOS	37.97	6.33	.21** [.10, .32]	.12* [.01, .24]	.14* [.02, .25]
18. SACT	21.52	4.88	.55** [.46, .64]	.43** [.32, .53]	.48** [.37, .57]

Note. M and *SD* are used to represent mean and standard deviation, respectively. MAT, SILS-V, and SILS-A refer to Miller Analogy Test practice items, Shipley Institute of Living Scale Verbal subscale, and Shipley Institute of Living Scale Abstraction subscale respectively. INF and IR refer to infrequency and instructed response respectively. SOS refers to student opinion scale. PSYN refers to psychometric synonyms. Values in square brackets indicate the 95% confidence interval for each correlation * indicates $p < .05$. ** indicates $p < .01$.

Table 5

Performance of RTE and Student Opinion Scale within Nomological Network

Nomological Network Component	MAT/SILS-V Response Time Effort	Student Opinion Scale
Correlation with SOS or MAT/SILS-V RTE	.18*/.06	.18*/.06
Moderation of SACT-MAT-V/SACT-SILS-V relationship	ΔR^2 =. 9%*/2.3%**	NS/NS
Moderation of SGPA-MAT-V/SGPA-SILS-V relationship	ΔR^2=2.1%**/4.4%**	NS/NS
Convergence with test performance on MAT/SILS-V	r = .46**/.31**	r = .22**/.14*
Discriminant validity with SACT	.13*/.09	.04
Warning effect	η^2 = .02**/NS	η^2 = .05**
Scale positioning effect	NS	NA

* p < .05. ** p < .01. NS = non-significant result. MAT = Miller Analogies Test practice items. SILS-V = Shipley Institute of Living Scale - Verbal subscale. SGPA = self-reported college grade point average. SACT = self-reported ACT scores. NA = non applicable.

Figure 1 - Plot of Significant Interaction Effect of MAT RTE on the Relationship between Self-Reported ACT scores and MAT Performance.

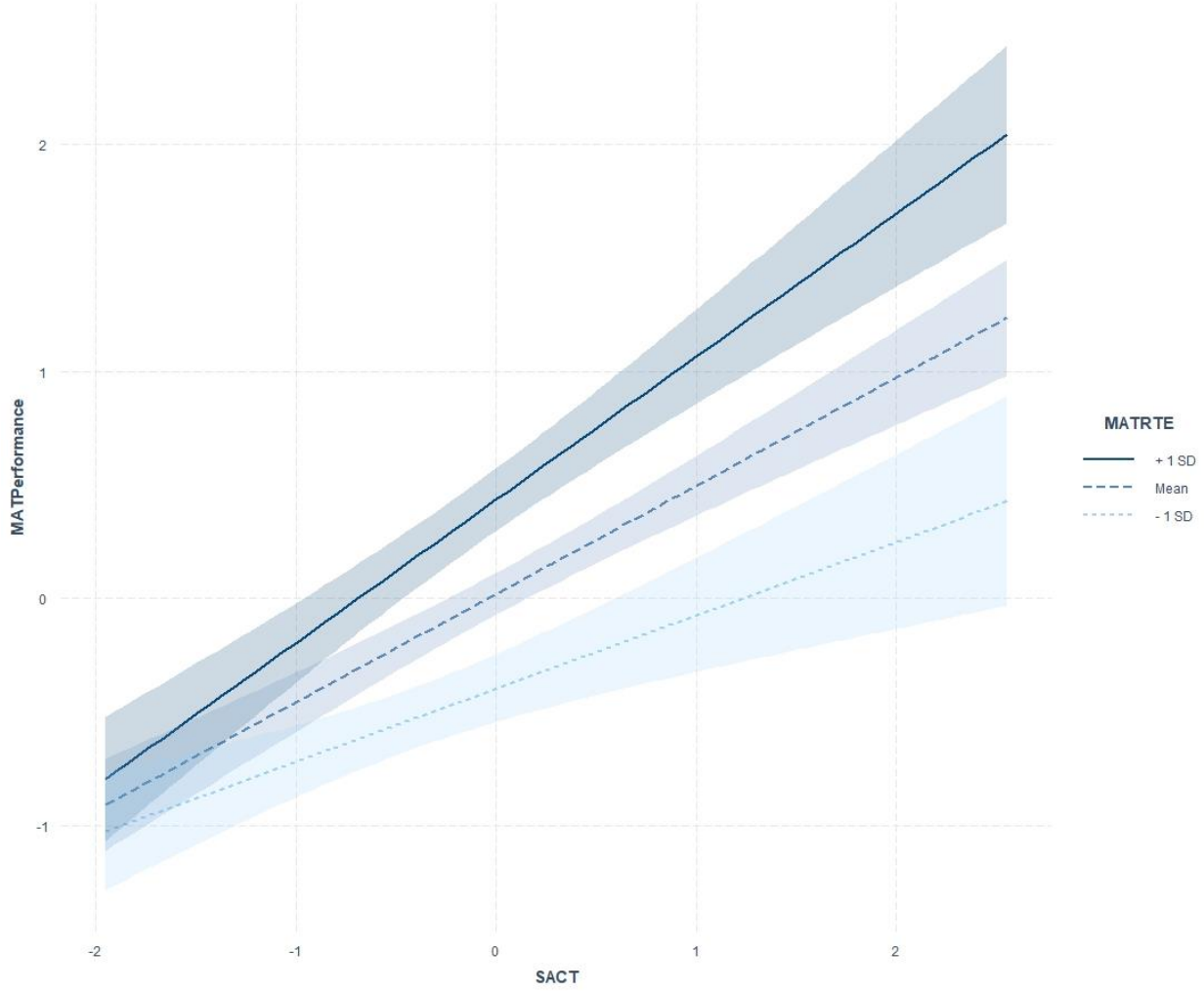

Figure 2 - Plot of Significant Interaction Effect of SILS-V RTE on the Relationship between Self-Reported ACT scores and MAT Performance.

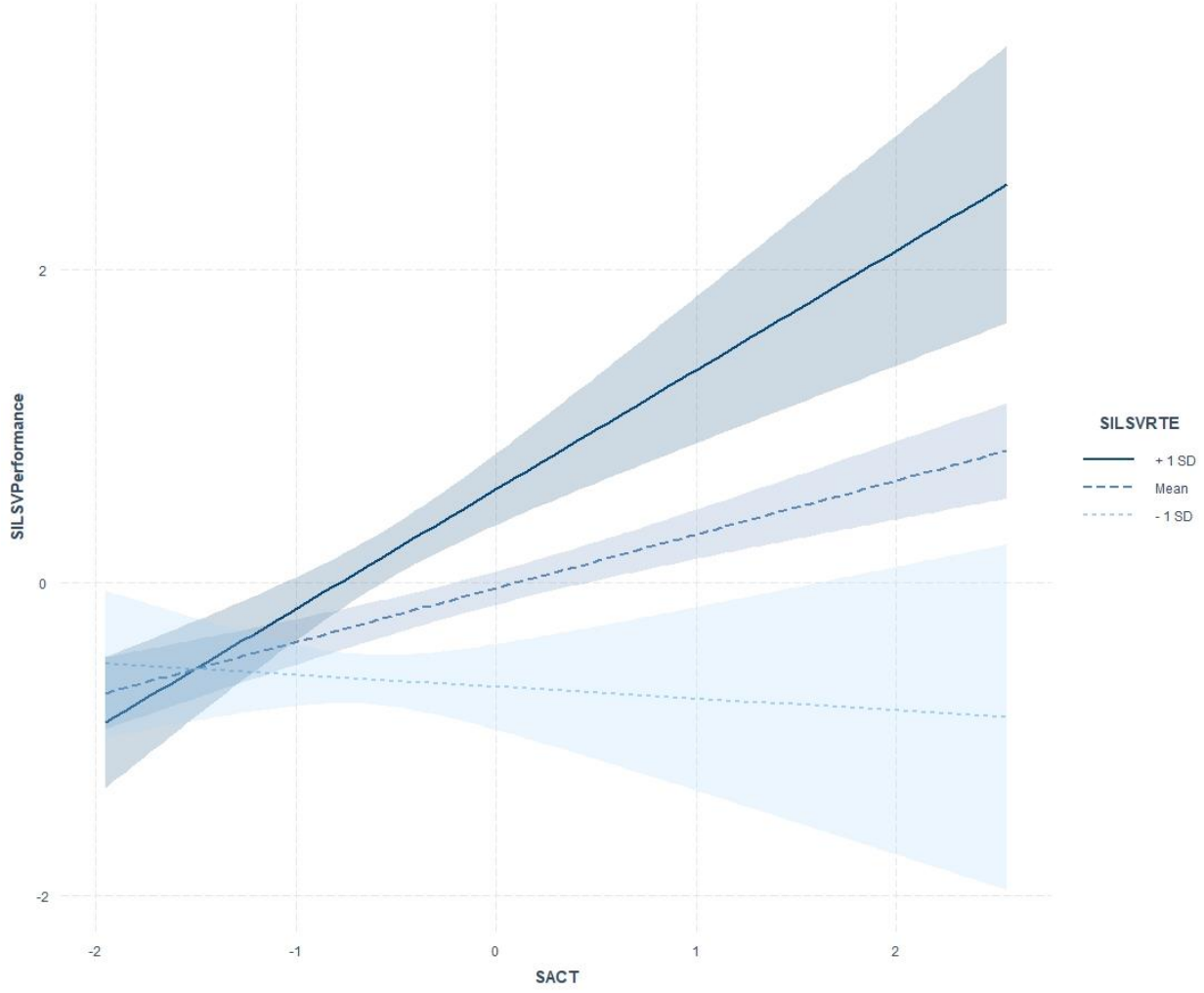

Figure 3 - Plot of Significant Interaction Effect of MAT RTE on the Relationship between Self-Reported college GPA and MAT Performance.

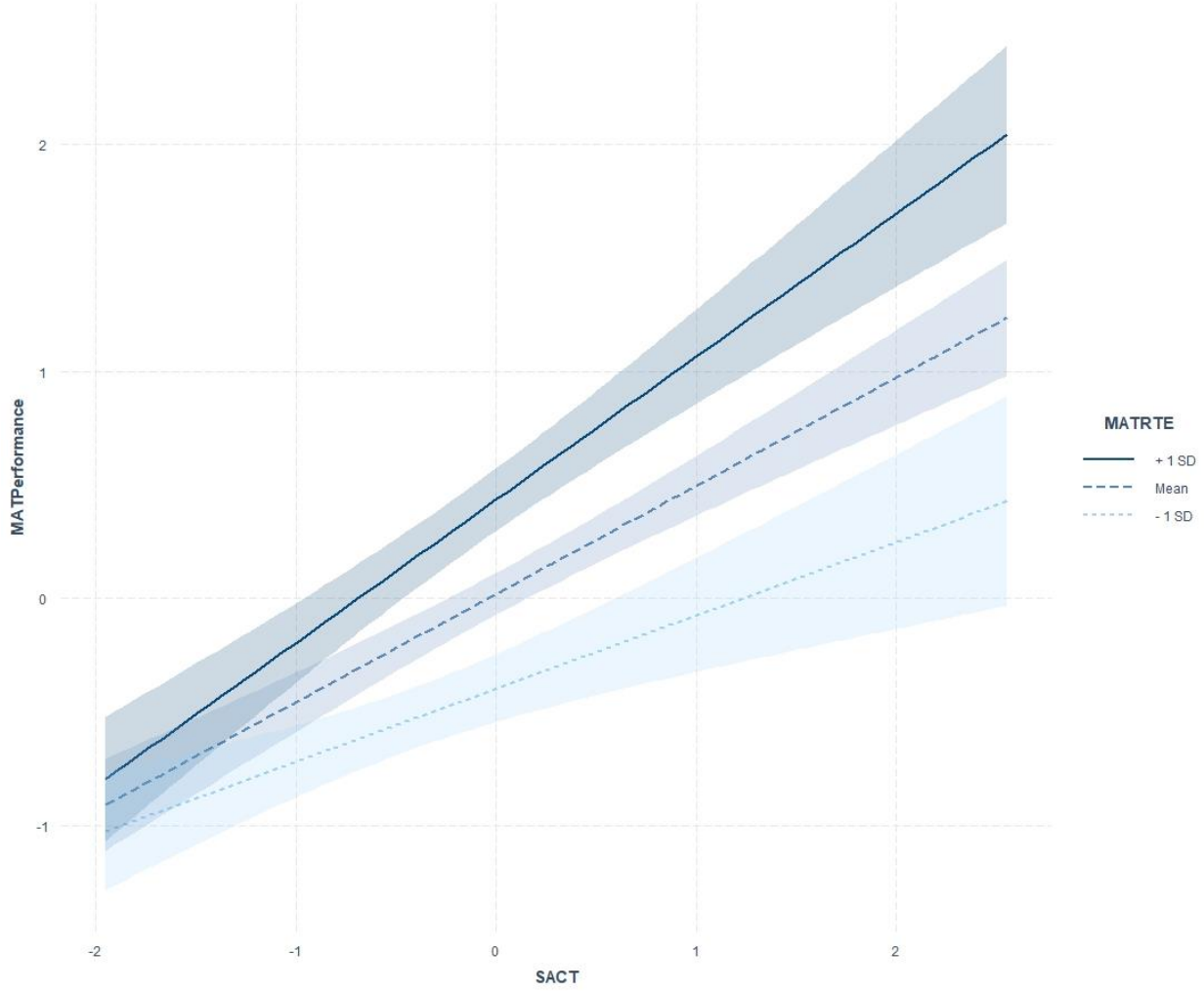

Figure 4 - Plot of Significant Interaction Effect of MAT Infrequency on the Relationship between Self-Reported college ACT scores and MAT Performance.

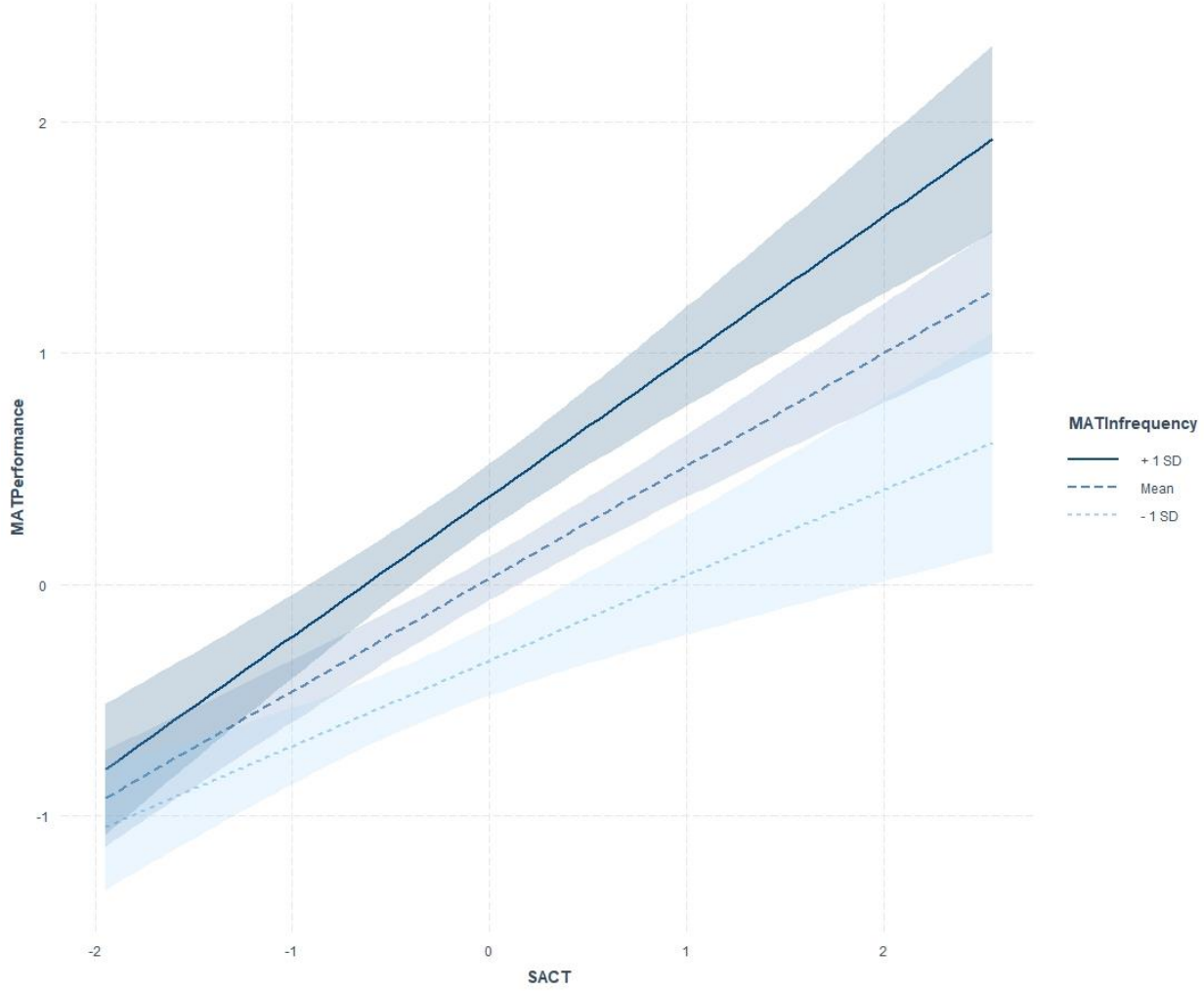

Figure 5 - Plot of Significant Interaction Effect of Long-String Total on the Relationship between Self-Reported college ACT scores and MAT Performance.

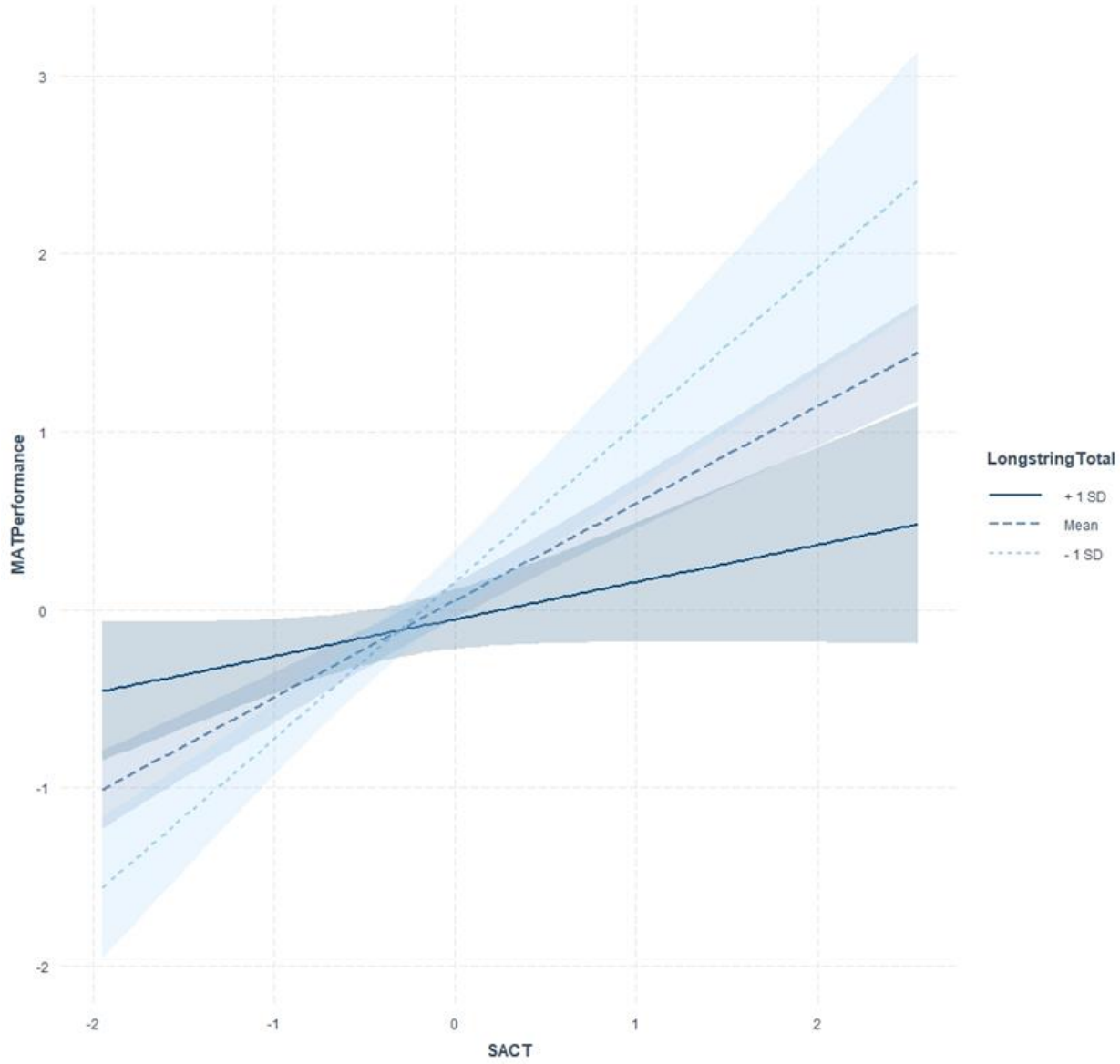

Figure 6 - Plot of Significant Interaction Effect of Long-String Total on the Relationship between Self-Reported college ACT scores and SILS-V Performance.

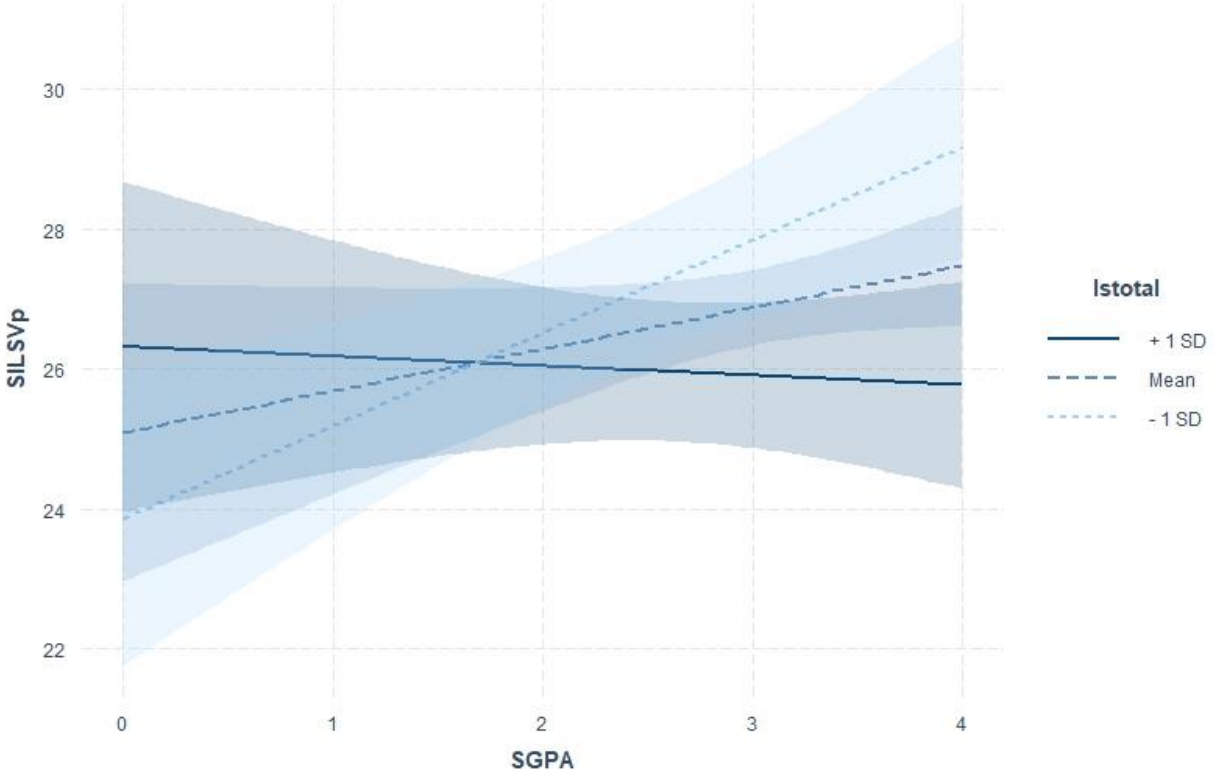

Table 6a1

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	-17.74**			
MATINF	3.39**	0.30		
SACT	1.18**	0.52		
			R^2 = .398**	
(Intercept)	-23.31**			
MATINF	4.11**	0.37		
SACT	1.14**	0.50		
MATINF x SACT	1.46*	0.12		
			R^2 = .409**	ΔR^2 = .011*

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *MAT* refers to Miller Analogies Test practice items. *MATINF* refers to Miller Analogies Test practice items infrequency items. SACT refers to self-reported ACT scores.

Table 6a2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	3.48			
SILS-V Infrequency	2.94**	0.28		
SACT	0.41**	0.42		
			R^2 = .264**	
(Intercept)	4.36			
SILS-V Infrequency	2.75**	0.26		
SACT	0.41**	0.42		
SILS-V Infrequency x SACT	-0.13	-0.02		

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *SILS-V* refers to Shipley Institute of Living Scale – Verbal subscale. SACT refers to self-reported ACT.

Table 6b

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	9.90**			
MATIR	-0.02	-0.00		
SACT	1.27**	0.55		
			R^2 = .307**	
(Intercept)	4.37			
MATIR	2.62	0.38		
SACT	1.52**	0.67		
MATIR x SACT	-0.12	-0.41		
			R^2 = .314**	ΔR^2 = .007

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *MAT* reflects Miller Analogies Test Practice Items. *MATIR* represents MAT instructed-response items. *SACT* refers to self-reported ACT scores.

Table 6b2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	6.91*			
SILS-VIR	3.77**	0.21		
SACT	0.40**	0.42		
			R^2 = .227**	
(Intercept)	-0.57			
SILS-VIR	6.31	0.35		
SACT	0.76	0.78		
SILS-VIR x SACT	-0.12	-0.40		
			R^2 = .229**	ΔR^2 = .001

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *SILS-V* reflects Shipley Institute of Living Scale – Verbal subscale. *SILS-VIR* represents SILS-V instructed-response items. *SACT* refers to self-reported ACT scores.

Table 6c

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	11.11**			
SACT	1.21**	0.54		
PS	-0.46	-0.01		
			R^2 = .289**	
(Intercept)	9.34			
SACT	1.28**	0.57		
PS	2.83	0.07		
PS x SACT	-0.14	-0.09		
			R^2 = .290**	ΔR^2 = .000

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates $p < .05$. ** indicates $p < .01$.

Table 6c2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	17.97**			
SACT	0.41**	0.41		
PS	-0.10	-0.01		
			R^2 = .173**	
(Intercept)	14.68**			
SACT	0.54**	0.55		
PS	6.02	0.36		
PS x SACT	-0.26	-0.38		
			R^2 = .178**	ΔR^2 = .005

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates $p < .05$. ** indicates $p < .01$.

Table 6d

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	3.34			
SACT	1.25**	0.55		
Diligence	0.14	0.07		
			R^2 = .312**	
(Intercept)	8.97			
SACT	0.98	0.43		
Diligence	0.02	0.01		
Diligence x SACT	0.01	0.14		
			R^2 = .312**	ΔR^2 = .000

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. SACT refers to self-reported ACT. MAT refers to Miller Analogies Test Practice items. * indicates $p < .05$. ** indicates $p < .01$.

Table 6d2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	13.32**			
SACT	0.41**	0.42		
Diligence	0.09*	0.11		
			R^2 = .198**	
(Intercept)	16.57			
SACT	0.25	0.26		
Diligence	0.03	0.03		
Diligence x SACT	0.00	0.19		
			R^2 = .199**	ΔR^2 = .000

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. SACT refers to self-reported ACT. SILS-V refers to Shipley Institute of Living – Verbal. * indicates $p < .05$. ** indicates $p < .01$.

Table 6e

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	9.95**			
LST	0.14	0.05		
SACT	1.26**	0.55		
			R^2 = .309**	
(Intercept)	9.69**			
LST	-0.37	-0.12		
SACT	1.28**	0.56		
LST x SACT	-3.79**	-0.24		
			R^2 = .340**	ΔR^2 = .031**

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. MAT refers to Miller Analogies Test Practice Items. LST refers to long-string total. SACT refers to self-reported ACT scores. * indicates $p < .05$. ** indicates $p < .01$.

Table 6e2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	17.79**			
LST	0.13	0.10		
SACT	0.41**	0.43		
			R^2 = .195**	
(Intercept)	17.73**			
lLST	-0.01	-0.01		
SACT	0.42**	0.43		
LST x SACT	-1.03	-0.16		
			R^2 = .207**	ΔR^2 = .012

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. SILS-V refers to Shipley Institute of Living Scale – Verbal Subscale. LST refers to long-string total. SACT refers to self-reported ACT. * indicates $p < .05$. ** indicates $p < .01$.

Table 7a

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	-13.96*			
MAT INF	4.47**	0.36		
SGPA	3.83**	0.20		
			R^2 = .184**	
(Intercept)	5.19			
MATINF	2.24	0.18		
SGPA	-2.49	-0.13		
MAT INF x SGPA	0.74	0.39		
			R^2 = .185**	ΔR^2 = .002

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *MAT* refers to Miller Analogies Test practice items. *MATINF* refers to Miller Analogies Test practice items infrequency items. SGPA refers to self-reported GPA.

Table 7a2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	11.85**			
SILS-VINF	2.33**	0.19		
SGPA	1.11*	0.14		
			R^2 = .062**	
(Intercept)	25.57*			
SILS-VINF	-0.47	-0.04		
SGPA	-4.68	-0.61		
SILS-VINF x SGPA	1.18	0.81		
			R^2 = .067**	ΔR^2 = .005

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *SILS-V* refers to Shipley Institute of Living Scale – Verbal subscale. *SILS-VINF* refers to Shipley Institute of Living Scale – Verbal subscale infrequency items. SGPA refers to self-reported GPA.

Table 7b

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	23.59**			
SGPA	4.27**	0.23		
MATIR	-0.06	-0.01		
			R^2 = .051**	
(Intercept)	28.46**			
SGPA	2.70	0.14		
MATIR	-2.24	-0.32		
MATIR x SGPA	0.70	0.33		
			R^2 = .055**	ΔR^2 = .004

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *SILS-V* reflects Shipley Institute of Living Scale – Verbal subscale. *SILS-VIR* represents SILS-V instructed-response items. *SGPA* refers to self-reported GPA.

Table 7b2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	27.67**			
SGPA	4.26**	0.23		
SILS-V IR	-1.42	-0.03		
			R^2 = .052**	
(Intercept)	39.19			
SGPA	0.63	0.03		
SILS-V IR	-5.31	-0.11		
SILS-VIRxSGPA	1.23	0.21		
			R^2 = .052**	ΔR^2 = .000

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates $p < .05$. ** indicates $p < .01$.

Table 7c

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	24.78**			
SGPA	4.41**	0.24		
PS	-4.19	-0.11		
			R^2 = .069**	
(Intercept)	19.43**			
SGPA	6.08**	0.33		
PS	6.36	0.16		
PS x SGPA	-3.30	-0.29		
			R^2 = .072**	ΔR^2 = .003

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates $p < .05$. ** indicates $p < .01$.

Table 7c2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	17.97**			
SACT	0.41**	0.41		
PS	-0.10	-0.01		
			R^2 = .173**	
(Intercept)	14.68**			
SACT	0.54**	0.55		
PS	6.02	0.36		
PS x SACT	-0.26	-0.38		
			R^2 = .178**	ΔR^2 = .005

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates $p < .05$. ** indicates $p < .01$.

Table 7d

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	8.44			
SGPA	4.05**	0.21		
Diligence	0.31**	0.16		
			R^2 = .078**	
(Intercept)	18.39			
SGPA	0.81	0.04		
Diligence	0.11	0.06		
Diligence x SGPA	0.06	0.21		
			R^2 = .079**	ΔR^2 = .000

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. SGPA refers to self-reported college grade point average. MAT refers to Miller Analogies Test Practice items. * indicates $p < .05$. ** indicates $p < .01$.

Table 7d2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	17.17**			
SGPA	1.17**	0.15		
Diligence	0.12**	0.15		
			R^2 = .050**	
(Intercept)	20.84			
SGPA	-0.02	-0.00		
Diligence	0.05	0.06		
Diligence x SGPA	0.02	0.19		
			R^2 = .051**	ΔR^2 = .000

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. SGPA refers to self-reported college grade point average. SILS-V refers to Shipley Institute of Living – Verbal. * indicates $p < .05$. ** indicates $p < .01$.

Table 7e

Regression results using MAT Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	30.91**			
LSTl	0.12	0.04		
SGPA	1.95*	0.14		
			R^2 = .022*	
(Intercept)	31.26**			
LST	-0.26	-0.08		
SGPA	1.87*	0.13		
LST x SGPA	-0.56	-0.14		
			R^2 = .029*	ΔR^2 = .007

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. MAT refers to Miller Analogies Test Practice Items. LST refers to long-string total. SGPA refers to self-reported college grade point average. * indicates $p < .05$. ** indicates $p < .01$.

Table 7e2

Regression results using SILS-V Performance as the criterion

Predictor	*b*	*beta*	Fit	Difference
(Intercept)	24.73**			
lstotal	0.12	0.09		
SGPA	0.68	0.11		
			R^2 = .025*	
(Intercept)	25.09**			
lstotal	-0.28*	-0.20		
SGPA	0.60	0.10		
lstxSGPA	-0.59**	-0.36		
			R^2 = .069**	ΔR^2 = .044**

Note. A significant *b*-weight indicates the beta-weight is also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. SILS-V refers to Shipley Institute of Living Scale – Verbal Subscale. LST refers to long-string total. SGPA refers to self-reported college grade point average. * indicates $p < .05$. ** indicates $p < .01$.

Table 8a

Fixed-Effects ANOVA Results Using MAT Infrequency Items as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	10234.35	1	10234.35	9022.17	.000		
Warning	1.83	1	1.83	1.61	.205	.01	[.00, .03]
Error	327.83	289	1.13				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 8a2

Fixed-Effects ANOVA Results Using SILS-V Infrequency Items as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90% CI [LL, UL]
(Intercept)	3390.86	1	3390.86	17494.12	.000		
Warning	0.00	1	0.00	0.02	.884	.00	[.00, .01]
Error	56.02	289	0.19				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 8b

Fixed-Effects ANOVA Results Using MAT Instructed-Response Items as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90% CI [LL, UL]
(Intercept)	749.83	1	749.83	280.64	.000		
Warning	7.18	1	7.18	2.69	.102	.01	[.00, .04]
Error	772.17	289	2.67				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 8b2

Fixed-Effects ANOVA Results Using SILS-V Instructed-Response Items as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	1218.35	1	1218.35	17615.95	.000		
Warning	0.02	1	0.02	0.28	.599	.00	[.00, .02]
Error	19.99	289	0.07				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 8c

Fixed-Effects ANOVA Results Using Psychometric Synonyms as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\,\eta^2$	$_{\text{partial}}\,\eta^2$ 90% CI [LL, UL]
(Intercept)	34.56	1	34.56	429.97	.000		
Warning	0.25	1	0.25	3.11	.079	.01	[.00, .04]
Error	22.67	282	0.08				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 8d

Fixed-Effects ANOVA results using Long-String Total as the criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	3.01	1	3.01	0.24	.625		
Warning	5.81	1	5.81	0.46	.498	.00	[.00, .02]
Error	3638.19	289	12.59				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 8e

Fixed-Effects ANOVA Results Using Diligence as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	333987.46	1	333987.46	9234.80	.000		
Warning	466.70	1	466.70	12.90	.000	.04	[.01, .09]
Error	10452.03	289	36.17				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 9a

Fixed-Effects ANOVA Results Using SILS-V Infrequency Items as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	3586.32	1	3586.32	18514.45	.000		
Scale Positioning	0.04	1	0.04	0.21	.649	.00	[.00, .01]
Error	55.98	289	0.19				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 9b

Fixed-Effects ANOVA Results Using SILS-V Instructed-Response Items as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	1299.33	1	1299.33	18806.41	.000		
Scale Positioning	0.04	1	0.04	0.58	.448	.00	[.00, .02]
Error	19.97	289	0.07				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 9c

Fixed-Effects ANOVA Results Using Psychometric Synonyms as the Criterion.

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\ \eta^2$	$_{\text{partial}}\ \eta^2$ 90% CI [LL, UL]
(Intercept)	39.35	1	39.35	485.41	.000		
Scale Positioning	0.06	1	0.06	0.71	.399	.00	[.00, .02]
Error	22.86	282	0.08				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 9d

Fixed-Effects ANOVA Results Using Diligence as the Criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90% CI [LL, UL]
(Intercept)	375706.87	1	375706.87	9945.21	.000		
Scale Positioning	0.98	1	0.98	0.03	.872	.00	[.00, .01]
Error	10917.75	289	37.78				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table 9e

Brown-Forsythe Test Results Using Long-String Total as the Criterion

Predictor	*df*	*F*	*p*
Scale Positioning	1	0.78	.375
Error	196.29		

Table 10 - Summary of Results Concerning Nomological Network

Nomological Network Component	Specific Test	Infrequency MAT/SILS-V	IR MAT/SILS-V	Psychometric Synonyms	LST	Diligence
Convergence with RTE	MAT RTE	.81	.10	-.02	.33	.15
	SILS-V RTE	.78	.31	-,09	.42	.05
Moderation of SACT-Test performance relationships	Moderation of SACT-MAT relationship	ΔR^2=1.1%	NS/NS	NS	NS	NS
	Moderation of SACT-SILS-V relationship	NS	NS	NS	NS	NS
Moderation of SGPA-Test performance relationships	Moderation of SGPA-MAT relationship	NS	NS	NS	NS	NS
	Moderation of SGPA-SILS-V relationship	NS	NS	NS	NS	NS
Convergence with test performance	MAT performance	.40	.02	-.11	.06	.15
	SILS-V performance	.36	.19	-.13	.10	.12
	SILS-A performance	.41/.15	.12/.19	-.28	.12	.13
Discriminant validity with SACT		.12/.04	.01/.07	-.15	.04	.04

Correlations < |.10| are nonsignificant (p > .05). NS refers to nonsignificant or that the effect was not significant in the predicted way. MAT = Miller Analogies Test practice items. SILS-V refers to Shipley Institute of Living Scale - Verbal subscale. SILS-A refers to Shipley Institute of Living Scale - Abstraction subscale. SGPA = self-reported college grade point average. SACT = self-reported ACT scores.

Table 12
Positive Participant Reactions to the Study Provided on Open-Response Question

Participant responses
"I personally think the study is super interesting and am glad I was able to participate."
"Good study, hope you have fun."
"I really enjoyed this study. Even though I could see where it was going I was a little surprised
"This [was a] very interesting test. I would like to know how I did on this test! Thank you for the opportunity."
Interesting study, this was the first one that I have done…"
"Very good study…"
"No complaints or anything, I just wanted to say that I find this study to be really interesting…"

Table 13

Proportion of Correct Responses to Infrequency Items

Item Stem	Proportion of Correct Responses
hungry : eat :: tired : _____	.96
kitten : cat :: puppy : _____	.94
eyes : sight :: nose : _____	.96
mountain : climb :: _____ : swim	.95
moon : night :: _____ : day	.98
plane : _____ :: boat : water	.94
banana : yellow :: broccoli : _____	.99
_____ : crawl :: adult : walk	.97
hot : _____ :: up : down	.96
arrow : bow :: pen : _____	.97
BABY	.98
FAST	.99
LARGE	.98
HOT	.98
TIRED	.99
ANGRY	.98
LOUD	.99

Table 14

Proportion of Correct Responses to Instructed-Response Items

Item Stem	Proportion of Correct Responses
Please select acorn	.98
Please select rain	.99
Please select variety	.98
_____ : please :: select : armor	.58
please : select :: visor : _____	.52
please : select :: batter : _____	.52
please : select :: sheath : _____	.53

APPENDIX A

Warning Message

PLEASE READ THE FOLLOWING INFORMATION CAREFULLY: It is vital to our study that participants devote their full attention to this survey. Otherwise years of effort (the researchers' time and the time of other participants) could be wasted. Please be aware that at the end of this survey, we will ask you to complete a multiple-choice quiz. This quiz will assess your knowledge of the content of the questionnaire and will be used to determine whether you have been paying attention. IF YOU DO NOT PASS THIS QUIZ, YOU MIGHT NOT RECEIVE COURSE CREDIT FOR COMPLETING THE SURVEY. Please answer the following questions to demonstrate that you read and understood the previous information.

Manipulation Check Items

Warning Manipulation Check Items

The researcher has told me that he or she will use advanced statistical techniques to detect the accuracy and thoughtfulness of my responses to today's questions
The researcher told me that I will lose my research credits if I fail to provide accurate and thoughtful responses to today's survey questions.

Note. Administered on a 7-point Likert scale ranging from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*).

APPENDIX B

Self-Report Diligence Scale

Items Included in Study

1. I carefully read every test item.
2. I could've paid closer attention to the items than I did.
3. I probably should have been more careful during these tests..
4. I worked to the best of my abilities in this study.
5. I put forth my best effort in responding to these tests..
6. I didn't give these tests the time it deserved.
7. I was dishonest on some items.
8. I was actively involved in this study.
9. I rushed through these tests.

APPENDIX C

Student Opinion Scale

1. Doing well on these tests was important to me.

2. I engaged in good effort throughout these tests.

3. I am not curious about how I did on these tests relative to otherS

4. I am not concerned about the scores I receive on these tests.

5. These were important tests to me.

6. I gave my best effort on these tests.

7. While taking these examinations, I could have worked harder on them.

9. I did not give these tests my full attention while completing them.

10. While taking these tests, I was able to persist to completion of the tasks.

APPENDIX D

Shipley Institute of Living Scale

Note* two subscales – verbal and abstraction. Abstraction has 25 items and 30 seconds per each item. Verbal has 50 items and 15 seconds per each item. 25 minutes total across the two tests.

ABSTRACTION TEST
Complete the following. Each dash (-) calls for either a number or a letter to be filled in. You will have 30 seconds to complete each item.

Stem
1 2 3 4 5 -
White black short long down - -
AB DB CD D -
Z Y X W V U -
1 2 3 2 1 2 3 4 3 2 3 4 5 4 3 4 5 6 - -
NE/SW SE/NW E/W N/-
Escape scape cape - - -
Oh ho rat tar mood - - - -
A Z B Y C X D -
Tot tot bard drab 537 - - -
Mist is wasp as pint in tone - -
Knit in spud up both to stay - -
Scotland landscape scapegoat - - - - ee

Surgeon 1234567 snore 17635 rogue - - - - -

Tam ran rib rid rat raw hip - - -

Tar pitch throw saloon bar rod fee tip end plank - - - - - meals

3124 82 73 154 46 13 -

Lag leg pen pin big bog rob - - -

Two w four r one o three -

Note.

APPENDIX E

SILS-Verbal items

Stem	Response Option 1	Response Option 2	Response Option 3	Response Option 4
TALK	draw	eat	speak	sleep
PERMIT	allow	sew	cut	drive
PARDON	forgive	pound	divide	tell
COUCH	pin	eraser	sofa	glass
REMEMBER	swim	recall	number	defy
TUMBLE	drink	dress	fall	think
HIDEOUS	sirvery	tilted	young	dreadful
CORDIAL	swift	muddy	leafy	hearty
EVIDENT	green	obvious	sceptical	Afraid
IMPOSTER	conductor	officer	book	Pretender
MERIT	deserve	distrust	fight	Separate
FASCINATE	welcome	fix	stir	Enchant
INDICATE	defy	excite	Signify	Bicker
IGNORANT	red	sharp	Uninformed	Precise
FORTIFY	submerge	Strengthen	Vent	Deaden
RENOWN	length	Head	Fame	Loyalty
NARRATIVE	yield	Buy	Associate	Tell
MASSIVE	bright	Large	Speedy	Low
HILARITY	lauehter	Speed	Grace	Malice
SMIRCHED	stolen	Pointed	Remade	Soiled

SQUANDER	tease	Belittle	Cut	Waste
CAPTION	drum	Ballast	Heading	Ape
FACILITATE	help	Turn	Strip	Bewilder
JOCOSE	humorous	Paltry	Fervid	Plain
APPRISE	reduce	Strew	Inform	Delight
RUE	eat	Lament	Dominate	Cure
DENIZEN	senator	Inhabitant	Fish	Atom
DIVEST	dispossess	Intrude	Rally	Pledge
AMULET	charm	Orphan	Dingo	Pound
INEXORABLE	untidy	Involatile	Rigid	Sparse
SERRATED	dried	Notched	Armed	Blunt
LISSOM	moldy	Loose	Supple	Convex
MOLLIFY	mitigate	Direct	Pertain	Abuse
PLAGIARIZE	appropriate	Intend	Revoke	Maintain
ORIFICE	Brush	Hole	Building	Lute
QUERULOUS	Maniacal	Curious	Devout	Complaining
PARIAH	Outcast	Priest	Lentil	Locket
ABET	Waken	Ensue	Incite	Placate
TEMERITY	Rashness	Timidity	Desire	Kindness
PRISTINE	Vain	Sound	First	Level

APPENDIX F

MAT Practice Items

Miller Analogies Test (MAT) Practice Items

Note 80 MAT practice items. 30 seconds per item. 40 minutes total. Items are from citation below.

Dermott, B., Gade, S., McLean, K., Recco, W., & Schultz, C. (2002). *501 Word Analogy Questions* (1st ed.). LearningExpress, LLC.

MAT Instructions:

Each of the following questions will present three words, you will need to provide a fourth word to go together with the other three to complete the analogy. For example, you might be given these three words.

Warm : hot :: _____ : hilarious

a. humid
b. raucous
c. summer
d. amusing

This item should be read as “warm is to hot as _____ is to hilarious”.

The solution should be amusing because, just as warm is a lesser amount of heat than hot, amusing is a lesser amount of enjoyment than hilarious. The amusing answer completes the analogy being made.

There is a timer at the bottom of the page, it will tell you how much time is left for each problem. You will have 30 seconds for each question.

_____ : trail :: grain : grail
a. train
b. path
c. wheat

d. Holy

particular : fussy :: _____ : subservient
a. meek
b. above
c. cranky
d. uptight

_____ : horse :: board : train
a. stable
b. shoe
c. ride
d. mount

tureen : _____ :: goblet : wine
a. napkin
b. soup
c. spoon
d. pilsner

son : nuclear :: _____ : extended
a. father
b. mother
c. cousin
d. daughters

coif : hair :: _____ : musical
a. shower
b. close
c. praise
d. score

7 8. feta : Greek :: provolone : _____
a. salad
b. Swiss
c. blue
d. Italian

18 9. moccasin : snake :: _____ : shoe
a. alligator
b. waders

c. asp
d. loafer

53 10. _____ : zenith :: fear : composure
a. apex
b. heaven
c. heights
d. nadir

80 11. pill : bore :: core : _____
a. center
b. mug
c. bar
d. placebo

24 12. pilfer : steal :: _____ : equip
a. return
b. damage
c. exercise
d. furnish

56 13. native : aboriginal :: naïve : _____
a. learned
b. arid
c. unsophisticated
d. tribe

61 14. junket : _____ :: junk : trash
a. trounce
b. trip
c. refuse
d. trinket

65 15. _____ : festive :: funeral : somber
a. tension
b. soiree
c. eulogy
d. sari

67 16. fetish : fixation :: slight : _____
a. flirt

b. sloth
c. insult
d. confuse

11 17. hovel : dirty :: hub : _____
a. unseen
b. prideful
c. busy
d. shovel

15 18. bog : _____ :: slumber : sleep
a. dream
b. foray
c. marsh
d. night

55 19. _____ : segue :: throng : mass
a. subway
b. church
c. transition
d. line

35 20. ragtime : United States :: raga : _____
a. cloth
b. country
c. piano
d. India

58 21. miserly : cheap :: homogeneous : _____
a. extravagant
b. unkind
c. alike
d. Friendly

46 22. skew : gloomy :: slant : _____
a. glee
b. foible
c. desperate
d. Gloaming

27 23. eider : _____ :: cedar : tree
a. snow
b. plant
c. duck
d. pine

5 24. gerrymander : divide :: filibuster : _____
a. bend
b. punish
c. delay
d. rush

23 25. vapid : _____ :: rapid : swift
a. inspired
b. turgid
c. wet
d. insipid

12 26. denim : cotton :: _____ : flax
a. sheep
b. uniform
c. sweater
d. Linen

13 27. obscene : coarse :: obtuse : _____
a. subject
b. obstinate
c. obscure
d. stupid

32 28. diamond : baseball :: court : _____
a. poker
b. jury
c. grass
d. squash

52 29. quixotic : pragmatic :: murky : _____
a. rapid
b. cloudy
c. clear
d. friendly

19 30. smear : libel :: heed : _____
a. represent
b. doubt
c. consider
d. need

17 31. nymph : _____ :: seraphim : angel
a. maiden
b. sinner
c. candle
d. priest

20 32. poetry : rhyme :: philosophy : _____
a. imagery
b. music
c. bi-law
d. theory

59 33. jibe : praise : _____ : enlighten
a. jib
b. delude
c. worship
d. wed

50 34. marshal : prisoner :: principal : _____
a. teacher
b. president
c. doctrine
d. student

29 35. fecund : infertile :: _____ : fleet
a. rapid
b. slow
c. fertilizer
d. damp

64 36. mend : sewing :: edit : _____
a. darn
b. repair
c. manuscript

d. makeshift

26 37. abet : _____ :: alone :: lone
a. bet
b. loan
c. wager
d. single

71 39. piercing : _____ :: hushed : whisper
a. diamond
b. watch
c. siren
d. ears

73 40. segregate : unify :: repair : _____
a. approach
b. push
c. damage
d. outwit

38 41. congeal : solidify :: _____ : char
a. conceal
b. singe
c. evaporate
d. charge

69 42. _____ : marsupial :: monkey : primate
a. opossum
b. ape
c. honeybee
d. moose

60 43. principle : doctrine :: living : _____
a. will
b. dead
c. likelihood
d. livelihood

75 44. _____ : climb :: recession : withdrawal
a. ascent
b. absence

c. dollar
d. absorption

54 45. myopic : farsighted :: _____ : obscure
a. benevolent
b. famous
c. turgid
d. wasted

68 46. shallot : _____ :: scallop : mollusk
a. shark
b. muscle
c. dessert
d. onion

39 47. conjugate : pair :: partition : _____
a. divide
b. consecrate
c. parade
d. squelch

37 48. _____ : excerpt :: exercise : maneuver
a. exception
b. passage
c. routine
d. cause

70 49. alphabetical : _____ :: sequential : files
a. sort
b. part
c. list
d. order

77 50. tacit : implied :: _____ : inferior
a. shoddy
b. taciturn
c. forthright
d. superior

21 51. implement : rule :: _____ : verdict
a. propose

b. render
c. divide
d. teach

41 52. vaunt : boast :: skewer : _____
a. flaunt
b. criticize
c. prepare
d. avoid

16 53. gambol : _____ :: gamble : bet
a. skip
b. win
c. bat
d. worship

47 54. rotation : earth :: _____ : top
a. planet
b. spinning
c. sun
d. expanding

62 55. gall : vex :: hex : _____
a. fix
b. jinx
c. index
d. vixen

43 56. monarch : _____ ::
king : cobra
a. queen
b. butterfly
c. royal
d. venom

51 57. iota : jot :: _____ : type
a. one
b. ilk
c. tab
d. jet

44 58. _____ : subject :: veer : path
a. object
b. prove
c. math
d. digress

30 59. pan : _____ :: ban : judge
a. band
b. critic
c. author
d. lawyer

14 60. _____ : oyster :: paddy : rice
a. aphrodisiac
b. mollusk
c. bed
d. sandwich

40 61. cicada : _____ :: collie : canine
a. fruit
b. mineral
c. cat
d. insect

66 62. huckster : _____ :: gangster : crime
a. corn
b. trucking
c. policeman
d. advertising

1 63. _____ : bedrock :: cement : foundation
a. mica
b. water
c. lava
d. sand

72 64. dolorous : _____ :: sonorous : loud
a. woozy
b. weepy
c. dull
d. sleepy

49 65. lapidary : _____ :: dramaturge : plays
a. cows
b. gems
c. rabbits
d. movies

9 66. penurious : _____ :: deep : significant
a. generous
b. stingy
c. decrepit
d. cavernous

79 67. somnolent : nap :: truculent : _____
a. sleepwalker
b. journey
c. war
d. mood

10 68. nictitate : _____ :: expectorate : spit
a. wink
b. stomp
c. quit
d. smoke

22 69. cytology : _____ :: geology : rocks
a. cyclones
b. psychology
c. pharmacology
d. cells

57 70. proboscis : _____ :: abdomen : gut
a. prognosis
b. nose
c. ear
d. nausea

42 71. rein : horse :: control panel : _____
a. pilot
b. bit

c. plane
d. rider

34 72. Argentina : Brazil :: _____ : Iran
a. Canada
b. Iraq
c. Ireland
d. Mexico

48 73. _____ : play :: sing : anthem
a. act
b. scene
c. theater
d. field

31 74. mouse : _____ :: flash : camera
a. rat
b. computer
c. cord
d. dessert

78 75. cushion : sofa :: shelf : _____
a. ledge
b. bookcase
c. storage
d. frame

36 76. scrub : wash :: sob : _____
a. cry
b. water
c. sad
d. tease

33 77. moisten : _____ :: cool : freeze
a. water
b. soak

c. oven
d. grow
6 78. persimmon : _____ :: cottontail : rabbit
a. cinnamon
b. oven
c. badger
d. berry

3 79. stars : astronomy :: _____ : history
a. battles
b. eclipse
c. horse
d. autumn

76 80. _____ : unity :: dearth : scarcity
a. belief
b. death
c. cohesion
d. fear

63 96. egregious : bad :: _____ : small
a. minuscule
b. tall
c. wicked
d. cheap

98. lawless : order :: captive : _____
a. trouble
b. punishment
c. jail
d. freedom

APPENDIX G

MAT Infrequency Items

For the item stems, I used analogies that would be easily understood. For the responses, I surrounded the correct response **(in bold)** with 3 completely unrelated responses to minimize difficulty.

kitten : cat :: puppy : _____

- **Dog**
- Commerce
- Serious
- Straw

Moon : night :: _____ : day

- Branch
- **Sun**
- Forum
- Hotel

_____ : crawl :: adult : walk

- Estate
- Paper
- **Baby**
- Erosion

Eyes : sight :: nose : _____

- **Smell**
- Actor
- Needle
- Sextant

Mountain : climb :: _____ : swim

- Secretary
- **Lake**
- Supper

- Awaken

Plane : _____ :: boat : water
- Engraving
- Collecting
- **Air**
- Insect

Banana : yellow :: Broccoli : _____
- Coil
- Jargon
- **Green**
- Miser

Arrow : bow :: pen : _____
- Observation
- Frame
- French
- **Paper**

Hungry : eat :: tired : _____
- **Sleep**
- Territory
- Crush
- Window

Hot : _____ :: up : down
- Bath
- Actor
- Friend
- **Cold**

APPENDIX H

MAT Instructed-Response Items

MAT instructed response items

Here are 4 items that have the instructions embedded within the item stem. To make these items, I took unused MAT practice items and altered the stems.

_____ : please :: select : armor

- Armor
- Belt
- Tyne
- Shoe

Please : select :: _____ : visor

- Button
- Visor
- Pullover
- Hood

Please : select :: batter : _____

- Griddle
- Cake
- Batter
- Oven

Please : select :: _____ : sheath

- Weapon
- Rifle
- Sheath
- Club

APPENDIX I

SILS-V Infrequency Items

Stem	Response Option 1	Response Option 2	Response Option 3	Response Option 4	Response Option 5
Baby	**Child**	Article	Lecture	Army	Productive
Large	Agony	**Big**	Wedding	Cruel	Thesis
Hot	Bond	Courage	**Warm**	Swim	Weigh
Loud	Stomach	Election	Count	**Noisy**	Cook
Angry	Baseball	Class	Instal	Slice	**Mad**
Fast	**Quick**	Shed	Gaffe	Fireplace	Greeting
Tired	Roar	Control	**Sleepy**	Engagement	Prescription

Note. Correct answer in bold.

APPENDIX J

SILS-V Instructed-Response Items

Stem	Response Option 1	Response Option 2	Response Option 3	Response Option 4	Response Option 5
Please select acorn	**Acorn**	Behavior	Bold	Pity	Precedent
Please select rain	Lemon	**Rain**	Like	Heal	Gloom
Please select variety	Feather	Integrated	Feed	Fist	**Variety**

Note. Correct answer in bold.

APPENDIX K

SILS-A Instructed-Response Items

Stem	Correct Answer
Please Enter 2 -	2
Please Enter 10 -	10

Note. Not included due to a clerical error.

APPENDIX M

SILS-A Infrequency Items

Stem	Correct Answer
A B C D -	E
5 4 3 2 -	1
7 8 9 10 -	11

Note. Not included due to a clerical error.