9-18-2018

Identifying Naturally-occurring Direct Assessments of Social-emotional Competencies: The Promise and Limitations of Survey and Assessment Disengagement Metadata

James Soland
NWEA

Gema Zamarro
University of Arkansas, Fayetteville

Albert Cheng
Harvard University

Colin Hitt
University of Arkansas, Fayetteville

UNIVERSITY OF ARKANSAS | College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

Identifying Naturally-occurring Direct Assessments of Social-emotional Competencies: The Promise and Limitations of Survey and Assessment Disengagement Metadata

James Soland, Gema Zamarro, Albert Cheng, and Collin Hitt

September, 2018

EDRE Working Paper 2018-06

char·asse·in
THE CHARACTER ASSESSMENT INITIATIVE

Identifying Naturally-occurring Direct Assessments of Social-emotional Competencies: The Promise and Limitations of Survey and Assessment Disengagement Metadata

James Soland*
NWEA

Gema Zamarro
University of Arkansas

Albert Cheng
University of Arkansas

Collin Hitt
Southern Illinois University

First Version: May, 2018

*Corresponding author: James Soland, NWEA, 121 N.W. Everett Street, Portland, OR 97209, Ph. (503) 444-6449, jim.soland@nwea.org

Abstract

Social-emotional learning (SEL) is gaining increasing attention in education policy and practice due to evidence that related constructs are strongly associated with long-term academic achievement and attainment. However, the work of educators to support SEL is hampered by a lack of available, unbiased measures of related competencies. In this manuscript, we review a recent and growing body of literature suggesting that metadata captured when assessments are administered via computer can provide data on not only test engagement, but also SEL constructs. Implications of this new source of data for practice, policy, and research are discussed.

Keywords: Social-emotional learning; metadata; test disengagement; survey effort
JEL codes: C80, C83, C91

1. Introduction

Social-emotional learning (SEL) is an old concept that is gaining new traction in education practice and policy. SEL is a term that encapsulates a huge swath of research related to educational psychology. Psychological constructs associated with SEL often fall into broad categories like interpersonal, intrapersonal, and deep cognitive competencies (Soland, Stecher, & Hamilton, 2013), and include relatively new concepts like grit (Duckworth & Quinn, 2009) and growth mindset (Dweck, 2006). One reason for the renewed interest in SEL is a growing body of research providing evidence on the importance of social-emotional competencies (beyond the effect of cognitive ability) to long-term educational outcomes like high school graduation and workforce outcomes like earnings (Almlund, Duckworth, Heckman, & Kautz, 2011; Belfield et al., 2015; Dweck, Walton, & Cohen, 2011; Heckman & Vytlacil, 2001).

This interest has manifested itself in policy and practice. For example, the California Office to Reform Education (CORE) is a consortium of districts serving over one million students that banded together in 2010 to get a waiver of provisions of the No Child Left Behind Act of 2001. Their revised accountability system included measuring outcomes like academic self-management, growth mindset, self-efficacy, and social awareness scores (West, 2016). More recently, the Every Student Succeeds Act (ESSA) of 2015—the main policy mechanism for federal accountability and newest instantiation of the Elementary and Secondary Education Act—requires states to include non-academic indicators, which are often related to SEL, in their accountability plans.

However, as is so often the case, policies that encourage SEL development may be moving faster than the realities of educational data and assessment. Measuring social-emotional competencies is integral to fostering them: without measures, educators cannot assess the progress of students over time with much accuracy and policymakers cannot evaluate the impact

of programs designed to foster them (Duckworth & Yeager, 2015). Despite the importance of measuring related constructs, there is a shortage of high-quality measures (Duckworth & Yeager, 2015; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). This shortage tends to take three forms. First, for some difficult to measure constructs like creativity or innovation, there are few if any measures available supported by sufficient validity evidence to recommend their use in the classroom (Soland et al., 2013). Second, there may be available instruments, but they take the form of self-report measures like surveys that can suffer from biases that may undermine inferences educators wish to make based on them (Duckworth & Yeager, 2015; Kyllonen, 2012; Piedmont et al., 2000). Third, even when there are available measures, districts may face logistical, financial, or political barriers to administering the assessments, which means no SEL data are collected (Duckworth & Yeager, 2015; Soland et al., 2013).

Recent work (Hitt, 2015; Hitt, Trivitt, & Cheng, 2016; Soland, Jensen, Keys, Wolk, & Bi, 2018; Zamarro, Cheng, Shakeel, & Hitt, 2018; Zamarro, Hitt, & Mendez, 2016; Zamarro, Nichols, Duckworth, & D'Mello, 2017), in conjunction with several prior studies (Barry & Finney, 2016; Barry, Horst, Finney, Brown, & Kopp, 2010; Hernández & Hershaff, 2014), suggests a new source of SEL data that educators may be able to use to help safeguard against biased scores from more traditional measures and to supply a proxy for related constructs in the absence of data. Specifically, we have identified tasks that students complete during regular schooling that are not designed to be direct SEL assessments, but nonetheless capture information related to constructs like academic self-management and conscientiousness. Further, the tasks we highlight often involve more standardized activities than typically occur in the classroom. This feature may reduce the influence of factors irrelevant to the construct of interest that are a problem when using behaviors like attendance as a proxy for social-emotional

constructs. In some regards, these tasks are like the measurement equivalent of natural experiments in economics: they are not meant to be direct assessments (just as natural experiments are not designed to randomize study participants), yet variation in outcomes from these naturally occurring phenomena can provide meaningful information as if they were intended for those purposes.

The tasks we study are related to achievement tests or surveys (both referred to interchangeably as "assessments" throughout the manuscript) that students take during the school year, including those used to meet state and federal assessment requirements. Although these assessments are often not meant to measure SEL at all, taking an assessment like a math achievement test requires not only knowledge of the academic content, but also a willingness to engage with the questions and the ability to remain focused (Wise, 2015). Our findings suggest that metadata from an assessment, including how long students spend on items, whether they provide an answer to a question, and how idiosyncratically they select responses, have shown potential to help address shortages in more formal SEL measures by directly quantifying behaviors related to constructs like academic self-management and conscientiousness.

In this paper, we review current research on the benefits of measuring SEL using observable behaviors, provide two examples of evidence supporting the connection between assessment metadata and certain SEL constructs, and then discuss implications and limitations for educational stakeholders. By compiling and analyzing existing research on how assessment engagement metadata relate to SEL, we hope to begin accruing evidence to support a validity argument for uses of these data in a SEL space, as well as map out future research needs to support such arguments.

2. Measuring SEL Using Observable Student Behaviors

There is already a substantial precedent for using quantifications of observed student behaviors to generate useful data on SEL, especially constructs related to academic engagement. In particular, the early warning systems literature, which is devoted to identifying and supporting students who are at risk of dropping out of school, relies heavily on behavioral indicators of disengagement (Allensworth, 2013; Balfanz, Herzog, & Mac Iver, 2007). As Farrington et al. (2012) point out, "academic behaviors are the visible, outward signs that a student is engaged and putting forth effort to learn. Because they are observable behaviors, they are also relatively easy to describe, monitor, and measure" (p. 8). For example, students who are chronically absent, fail courses, and are suspended often are much more likely to drop out (Allensworth, 2013; Balfanz et al., 2007). These major disengagement behaviors often begin with much milder behaviors like coming to class unprepared and struggling to complete independent work (Farrington et al., 2012). Students who exhibit enough of these small behaviors often have more general issues with academic self-efficacy, self-management, conscientiousness, and grit, all of which are SEL constructs gaining prominence in research and policy due to their association with dropout (Bandura, 1994; Briesch & Chafouleas, 2009; Angela Lee Duckworth & Quinn, 2009; Zamarro et al., 2017). These constructs are defined in Table 1.

One potential problem with gaining data on SEL by observing real behaviors that occur during schooling is that such behaviors can often be related to a host of factors that have nothing to do with a particular social-emotional construct. For example, while students who fail courses may have low academic self-efficacy, they may also receive very low grades due to personal or situational issues unrelated to self-efficacy. Researchers have responded to these limitations by developing direct, performance-based assessments of social-emotional competencies (Miller & Linn, 2000). By constraining the conditions in which data are collected the hope is to help

standardize results and potentially remove irrelevant sources of variance. Such measures assess these competencies by having students directly perform tasks that relate to the construct of interest, which helps avoid self-report bias and may provide more authentic assessments of multifaceted constructs like creativity (Soland et al., 2013).

There are many examples of performance assessments being used to measure constructs related to SEL. A classic example is Mischel, Ebbesen, and Zeiss's (1972) famous "Marshmallow Test," which was designed to measure self-regulatory skills that are highly related to constructs like self-management. In more recent times, Galla et al. (2014) developed an Academic Diligence Task, a performance assessment that further standardizes the initial Marshmallow Test. One problem with such assessments is that construct-irrelevant variance can still be an issue if contextual factors influence results (Shavelson, Baxter, & Pine, 1991). A science performance task might produce biased results for a student if, say, a pipette breaks. To overcome such challenges, computer technology is being used to make contextual factors more standard (Soland et al., 2013). For example, the Programme for International Student Assessment (PISA) now offers a test of collaborative problem solving during which the student directly collaborates with an avatar, a simulated person with known problem-solving and teamwork capacities.

Despite advances in performance assessments that can help avoid self-report bias and standardize conditions in ways that can reduce construct-irrelevant bias, they still have limitations. First, tasks are generally very costly and difficult to collect in large samples, although new technologies are making this easier (Soland et al., 2013). Second, it is not always clear that artificial tasks completed in highly constrained settings are generalizable to other contexts (Bardsley, 2008; Duckworth & Yeager, 2015; Falk & Heckman, 2009). Finally, existing

performance tasks can be difficult to implement multiple times, as participants might gain familiarity after having performed the task once, upwardly biasing subsequent scores (Bardsley, 2008; Duckworth & Yeager, 2015; Falk & Heckman, 2009). Given these challenges, the pace at which performance tasks are developed is slow, and their adoption among educators may be even slower, further contributing to the shortage in available SEL measures (CASEL, 2006; Duckworth & Yeager, 2015; Soland et al., 2013).

3. Emerging Evidence on the Relationship between Metadata and Social-Emotional Competencies

Until recently, virtually no research considered the use of test and survey behavioral metadata to gain information on students' SEL needs. Using metadata is essentially a hybrid of observing student behaviors in school and measuring student behaviors in a controlled environment akin to those in performance tasks. While the metadata are captured during the administration of assessments that occur during the course of schooling, the conditions are often more consistent than during regular classroom instruction due to standardized protocols surrounding testing. While most assessments are not designed to capture behaviors related to SEL (e.g. skipping items on a survey), related metadata are often available.

We provide two broad examples of how assessment metadata are captured: one from achievement tests, the other from surveys. We discuss evidence showing a connection between these quantified assessment disengagement behaviors and SEL competencies related to academic self-management, self-efficacy, conscientiousness, and grit. Table 2 shows results from the studies we discuss related to achievement test metadata, including the authors, data sources, types of assessment metadata, and findings. Table 3 shows the same, but for survey metadata.

3.1 Metadata from Achievement Tests

While achievement test metadata have been used as a measurement tool for decades, those data are typically used to address measurement problems on those tests, not to provide information on social-emotional competencies. The most comprehensive work on achievement test metadata was developed by Wise (and catalogued in Wise, 2015). He and his colleagues showed that student engagement on achievement tests can be measured by identifying responses to items that are provided so rapidly, the content of those items could not have been understood (Demars, 2007; Rios, Liu, & Bridgeman, 2014; Wise & Kong, 2005). For example, if a student responds to an item with a lengthy reading passage in under 10 seconds, one can be fairly certain the student did not engage with that item. This behavior is often referred to as "rapid guessing" because students who respond rapidly enough get items correct at a rate no better than chance (Demars, 2007; Kong, Wise, & Bhola, 2007; Wise & Kong, 2005). Rapid guessing is largely uncorrelated with academic ability, meaning that this behavior is not just occurring because students do not understand the content (Wise, 2015).

Emerging research shows that the amount of time students spend on achievement test items—and whether students rapidly guessed—is related to more than test engagement. Work conducted by Soland, Jensen, Keys, Wolk, and Bi (2018) showed that rates of rapid guessing are related to social-emotional competencies, and to broader disengagement from school. Table 2 highlights relevant results from Soland et al. (2018). In terms of SEL, partial correlations between rapid guessing rates and self-management scores were 0.26, and the same correlations for self-efficacy were 0.12 (both significant at the .01 level). In terms of academic disengagement, which often stems from factors like low self-management and self-efficacy

(Farrington et al., 2012), Table 2 shows that students who rapidly guessed on 10% or more of the items on a given test had lower GPAs and attendance, as well as higher rates of suspensions and detentions. In tandem, these findings may suggest that students who rapidly guess are often disengaging from not only the test, but from school more generally, and may be at risk of dropping out.

Soland (2018) also found that rapid guessing can provide information on how students respond to academic challenge. Students in his sample were 18 times as likely to rapidly guess on difficult items if they were in the bottom quartile of self-efficacy scores compared to students in the top quartile. Similarly, students spent 1.5 times as long on very difficult items if their self-efficacy scores were in the top quartile rather than the bottom. Thus, achievement tests may capture data on how students respond to challenging tasks by capturing duration data that help quantify whether students persist on especially difficult items.

Other measures of test engagement related to SEL include measures of decline in performance as the test progresses, as well as the number of questions skipped on tests. Borghans and Schils (2012) computed measures of test fatigue, or how much students' performance declined throughout the testing period. They found that test fatigue was related to SEL factors such as motivation and conscientiousness and was predictive of educational attainment, employment, and earnings in adulthood. Beyond test fatigue, Hernández and Hershaff (2014) measured how often students skip questions on state standardized tests. They found that the skipping questions was associated with lower probabilities of high school graduation and college enrollment among students in Michigan.

3.2 Metadata on Surveys

Taking tests of academic achievement are not the only assessments that students do in school. Increasingly, students are also given surveys to, for example, assess school climate, evaluate their teachers, or disclose personal information about themselves. Like achievement tests, surveys require more than basic literacy skills and cognitive ability to complete them. They also require that students engage and exert effort to respond to each item (Curran, 2015; Meade & Craig, 2012). According to Curran et al. (2010), disengaged responding has been documented at rates ranging from 5% to 50% of collected surveys, depending on the context and detection method. In some cases, disengaged responding manifests itself when students skip survey items even when they have the requisite knowledge and understanding of the question to respond (Hitt et al., 2016). In other cases, students simply provide careless or inconsistent answers, such as when they repeatedly use only one response category on a Likert scale or select the same scale response category on two items measuring oppositional constructs, e.g. confidence in math and self-doubt in math (Hitt, 2015; Zamarro et al., 2018).

These two behaviors, which we will call "item nonresponse" and "careless answering," respectively, can be quantified and provide evidence on how engaged a student is on the survey. Item nonresponse rate is defined as the percentage of items skipped by a student out of the total number of items the student was supposed to answer in a survey (Hitt et al., 2016). Careless answering captures the prevalence of inconsistent answering on a survey for a student. Technical details for constructing this measure are described in Hitt (2015) and Zamarro et al. (2018). Intuitively, responses to items that are a part of a scale designed to measure a single construct should be correlated with each other. The careless answering measure captures the extent to which the responses are uncorrelated as in the case where a student always selects the first answer option even when doing so is logically inconsistent given the content of the survey.

While research suggests that inconsistent responding is not always a perfect proxy for test engagement, there is consistent evidence that such responses generally do not provide useful data on the construct being measured (Wise & Kong, 2005).

Both of these measures have been shown to capture information about SEL competencies like conscientiousness and grit. Table 3 summarizes the research evidence. Students with higher item nonresponse rates or careless answering scores self-report lower levels of grit and self-control (Zamarro et al., 2017). Partial correlations between these self-reported measures and measures of survey engagement are about 0.2. While the correlations are not high, they are comparable in magnitude to correlations among SEL survey scores, and between SEL scores and achievement, in other studies (Farrington et al., 2012; Gil-Olarte Marquez et al., 2006; Soland, Stecher, & Hamilton, 2013). This relationship between careless answering and constructs like self-management and grit were also observed in adulthood through an internet panel representative of American adults (Zamarro et al., 2018).

As with measures of disengagement from achievement tests, survey item nonresponse rates and careless answering were also found to be associated with later life outcomes like educational attainment, employment, and earnings, even after controlling for cognitive ability and demographic background characteristics (Hedengren & Stratmann, 2012; Zamarro et al., 2017).[1] Further, these correlations are not merely contemporaneous. Item nonresponse rates and careless answering in adolescence have both been found to predict these long-run life outcomes (Hitt, 2015; Hitt et al., 2016).

4. Potential Uses of Assessment Engagement Metadata to Support SEL Policy, Practice, and Research

[1] There is also evidence that the extreme case in which respondents fail to even begin a survey occurs more often among less conscientious respondents (Cheng et al., 2018; Lugtig, 2014).

In order to promote SEL, educators need to be able to measure related competencies. Without related data, educational stakeholders cannot tell if students' SEL competencies are improving, and whether programs to promote those competencies are working. Thus, establishing the relationship between assessment engagement metadata and SEL has several practical benefits. We discuss three of them.

First, such metadata can be used to help validate student scores from surveys (or other measures) of SEL competencies. Students are often unaware that computer-based assessments capture metadata like response times and proportions of omitted responses. Therefore, not only is self-report bias avoided, but there may also be a lower likelihood that students will behave differently due to awareness of the behavior being measured. A measure with these properties can prove useful to scrutinizing self-reported measures. For example, if a student reports high self-management or conscientiousness, but rapidly guesses frequently on an achievement test or omits responses on a survey, then educators might worry about self-report bias. One of the most novel facets of this multiple-measures approach is that metadata from a survey can serve as a check against self-report bias on that same survey (Duckworth & Yeager, 2015).

Second, assessment engagement metadata may also be useful to administrators and teachers by supplementing datasets that do not have SEL scores. For example, practitioners could benefit by gaining a proxy for certain SEL constructs if a district or school does not offer a survey (Soland et al., 2018). Even in the event a school system does measure SEL through a survey or other instrument, those measures are often administered no more than yearly. Thus, such districts could gain SEL data between survey administrations by relying on metadata from other assessments. Notably, this multiple-measures approach is extremely cheap because it does not require administration of an additional assessment, which means districts can get additional

SEL data from the testing regimes they already have in place. There may be similar benefits for researchers: many large publicly available datasets do not include scores from SEL measures despite the fact that social-emotional data might support useful research with the dataset (Zamarro et al., 2016).

Finally, assessment behavior metadata could provide early warning indicators that a student is at risk of academic disengagement. Low academic engagement is associated with reduced educational attainment, including failing to complete high school (Farrington et al., 2012). The early warning literature typically highlights other behaviors like suspensions or absenteeism when trying to identify disengaged students (Allensworth, 2013). Research shows that assessment disengagement behaviors are similar to behaviors in the early warning indicator literature suggesting academic disengagement (Hitt et al., 2016; Soland et al., 2018). Therefore, associated metadata may provide another behavioral early warning indicator of whether a student is academically disengaged and potentially at risk of dropping out. Given how often students are tested in schools currently, these metadata are captured quite frequently, which could also increase their value.

5. Potential Limitations of Assessment Engagement Metadata to Support SEL Policy, Practice, and Research

Despite the promise of using metadata in a multiple-measures approach to assessing SEL competencies, there are several major limitations. First, much more validity evidence would need to be collected to argue that these behavioral indicators are actually measures of constructs like self-management and, even then, there might be too many confounding factors. As one example, response time metadata can be impacted by constraints that schools or districts place on tests (e.g. when in the day they are administered), which could change behaviors in ways

irrelevant to the construct of interest (Wise, 2015). For another, students might be more likely to respond carelessly to assessment items if those questions are poorly worded (Curran, 2015). Although the emerging literature is promising, until more validity evidence is collected to support particular SEL-based uses of assessment disengagement metadata, one might be safer thinking of those metadata as crude proxies for SEL competencies like conscientiousness rather than as valid measures.

Second, there may not be straightforward ways to reconcile discrepant results from surveys and metadata. For instance, a student may report low self-management yet rapidly guess infrequently if at all. More still needs to be learned about cross-classification rates between measures. Put differently, assessment disengagement behavior may be insufficient on its own to establish issues with self-management or conscientiousness. At best, one would imagine that such metadata could be part of a multiple-measures approach to assessing SEL.

Finally, assessment disengagement metadata are not especially helpful in an accountability context because, like survey scores, they can be easily gamed. Even if educators and students did not know how assessment metadata were being used, exactly, a general awareness of test metadata being used for accountability could incent perverse activities. For instance, if behaviors paralleled what has been seen on achievement tests, educators might coach their students to spend long amounts of time on items, or even bubble in items their students left blank (Jones, 2011). While such responses to the inclusion of metadata in accountability systems may not occur, there is a strong argument to be made that assessment engagement metadata should be used primarily for low-stakes purposes among educators, policymakers, and researchers.

6. Conclusion and Future Research

There is increasing evidence that, to succeed in life, students need to leave school with more than knowledge of academic subject matter (Dweck et al., 2011). In this paper, we considered the promise and limitations of naturally occurring behaviors that provide data on certain social-emotional constructs. These behaviors are somewhat like a measurement version of natural experiments in economics, which are not designed to randomize students, but allow for related inferences anyway. We reviewed a growing body of research showing that metadata captured when students take achievement tests or surveys can provide insight not only into engagement on the assessment, but also to SEL constructs including self-management and conscientiousness (Hitt, 2015; Hitt et al., 2016; Soland, 2018a, 2018b; Zamarro et al., 2018, 2016, 2017). Studies suggest these assessment engagement metadata may be beneficial as a check against self-report bias on SEL surveys, to supplement SEL data used in practice when available data are sparse or nonexistent, and to serve as early warning indicators that a student may have begun to disengage academically. This literature review is meant to provide the foundation for a validity argument supporting these uses of metadata, which may be useful to educators as they try to foster SEL.

Going forward, each of these potential uses should be supported with additional validity evidence. For example, research should further explore how well assessment engagement metadata perform as early warning indicators of dropout relative to more established indicators like chronic absenteeism. Studies might also consider whether inferences about student progress and program effectiveness related to SEL are consistent when metadata are used versus self-report surveys. Such validation work would likely benefit from being conducted in concert with educators who use SEL data to support their practice on a regular basis.

References

Allensworth, E. (2013). The Use of Ninth-Grade Early Warning Indicators to Improve Chicago Schools. *Journal of Education for Students Placed at Risk (JESPAR)*, *18*(1), 68–83.

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality psychology and economics*. National Bureau of Economic Research.

Balfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*(4), 223–235.

Bandura, A. (1997). Self-efficacy: The exercise of control. New York, NY: Macmillan.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, *11*(2), 122–133.

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education*, *29*(1), 46–64.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363.

Belfield, C., Bowden, A. B., Klapp, A., Levin, H., Shand, R., & Zander, S. (2015). The economic value of social and emotional learning. *Journal of Benefit-Cost Analysis*, *6*(3), 508–544.

Borghans, L., & Schils, T. (2012). The leaning tower of PISA: The effect of test motivation on scores in the international student assessment. Paphos, Cyprus: Paper presented at the EALE annual conference, September 2011.

Briesch, A. M., & Chafouleas, S. M. (2009). Review and analysis of literature on self-management interventions to promote appropriate classroom behaviors (1988–2008). *School Psychology Quarterly*, *24*(2), 106.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor,R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. Child Development, 82, 405-432.

Cheng, A. (2015). Like Teacher, Like student: Teachers and the development of student noncognitive skills. Fayetteville: University of Arkansas.

Cheng, A., Zamarro, G., & Orriens, B. (2018). Personality as a predictor of unit nonresponse in an internet panel. Sociological Methods and Research.

Curran, P. G. (2015). 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.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

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

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432.

Dweck, C., Walton, G. M., & Cohen, G. L. (2011). Academic tenacity: Mindset and skills that promote long-term learning. *Gates Foundation. Seattle, WA: Bill & Melinda Gates Foundation*.

Dweck, C. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance--A Critical Literature Review*. Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.

Galla, B. M., Plummer, B. D., White, R. E., Meketon, D., D'Mello, S. K., & Duckworth, A. L. (2014). The Academic Diligence Task (ADT): Assessing individual differences in effort on tedious but important schoolwork. *Contemporary Educational Psychology*, *39*(4), 314–325.

Heckman, J., & Vytlacil, E. (2001). Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. *Review of Economics and Statistics*, *83*(1), 1–12.

Hedengren, D., & Stratmann, T. (2012). "The Dog that Didn't Bark: What Item Nonresponse Shows about Cognitive and Non-Cognitive Ability." Unpublished Manuscript. doi:10.2139/ssrn.2194373

Hernández, M., & Hershaff, J. (2014). Skipping questions in school exams: the role of socio-emotional skills on educational outcomes. *Draft Version: March*, *18*, 2014.

Hitt, C. E. (2016). Just Filling in the Bubbles: Using Careless Answers Patterns on Surveys as a

Proxy Measure of Noncognitive Skills, EDRE Working Paper 2015-6. Fayetteville, AR: Department of Education Reform, University of Arkansas.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119.

Jones, D. L. (2011). Academic dishonesty: Are more students cheating? *Business Communication Quarterly*, *74*(2), 141–150.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Kyllonen, P. C. (2012, May). Measurement of 21st century skills within the common core state standards. In *Invitational Research Symposium on Technology Enhanced Assessments* (pp. 7-8).

Lugtig, P. J. 2014. "Panel attrition: Separating stayers, fast Attriters, gradual attriters, and lurkers. Sociological Methods and Research, 43(4),699–723

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, *24*(4), 367–378.

Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, *21*(2), 204.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*(3), 582.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying Low-Effort Examinees on Student Learning Outcomes Assessment: A Comparison of Two Approaches. *New Directions for Institutional Research*, *2014*(161), 69–82.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347–362.

Soland, J. (2018). Can the Amount of Time Students Spend on Test Items Help Illuminate Social-emotional Learning Needs? Initial Evidence from an International Achievement Test. Presented at the American Education Finance and Policy (AEFP), Portland, OR.

Soland, J., Jensen, N., Keys, T., Wolk, E., & Bi, S. (2018). *Is Low Test Motivation a Sign of Disengagement from School? Examining Indicators of Dropout Conditional on Rapid Guessing Scores.* Manuscript submitted for publication.

Soland, J., Hamilton, L.S, & Stecher, B. M. (2013). Measuring 21st Century Competencies. Santa Monica, CA: RAND Corporation.

West, M. R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports*, *1*(13).

Wise, S. L. (2015). Effort Analysis: Individual Score Validation of Achievement Test Data. *Applied Measurement in Education*, *28*(3), 237–252.

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

Zamarro, G., Cheng, A., Shakeel, M. D., & Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics*, *72*, 51–60.

Zamarro, G., Hitt, C., & Mendez, I. (2016). When Students Don't Care: Reexamining International Differences in Achievement and Non-Cognitive Skills. EDRE Working Paper No.2016-18. Fayetteville, AR: Department of Education Reform, University of Arkansas.

Zamarro, G., Nichols, M., Duckworth, A., & D'Mello, S. (2017). Further validation of survey effort measures of conscientiousness: results from a sample of high school students. Washington DC, USA: Paper presented at the 42nd AEFP annual conference, March 2017.

Table 1

Definition of Constructs Used in this Study

Social-emotional Construct	Definition	Citation
Academic Self-efficacy	Self-efficacy is a student's self-confidence in his or her academic abilities, and is a fundamental building block of motivation in school. If students do not believe they can complete an academic task, then they have little incentive to undertake it.	(Bandura, 1993)
Academic Self-management	Self-management is students' ability to focus on academic tasks and regulate their own academic behavior. Students with low self-management are much more likely to fail courses and have lower attendance, both of which are associated with high school dropout.	(Briesch & Chafouleas, 2009)
Conscientiousness	The tendency to be organized, responsible, and hardworking.	(American Psychology Association Dictionary)
Grit	Trait-level perseverance and passion for long-term goals.	Duckworth & Quinn, 2009)

Table 2

Summary of Findings on The Relationships among Achievement Test Metadata, SEL, and Academic Engagement

Study	Data	Metadata Source	Findings	
			SEL	Academic Engagement
Barry & Finney (2016)	University sophomores and juniors completing a low-stakes, three-hour testing session	Test Engagement Surveys, Latent Growth Modeling Estimates	Changes in test engagement over the course of the test related to agreeableness, conscientiousnes	
Barry, Horst, Finney, Brown, & Kopp (2010)	Incoming first-year students who completed a three-hour testing session during a university-wide assessment day at a mid-sized southeastern U.S. university	Item Durations	Test engagement was correlated with Big 5 personality characteristics	
Borghans & Schils (2011)	Dutch Inventaar 2010 data	Decline in Test Effort	Declines in test effort are greater among students with lower levels of grit and conscientiousness.	Declines in test effort are greater among students who report lower levels of motivation to go to school and motivation to learn. Parents report higher rates of absence from school for students with greater levels of test decline.
Hernandez & Hershaff (2014)	Longitudinal data from the Michigan Student Data Systems	Item Nonresponse		Skipping multiple questions in one of the 7th or 8th grade standardized tests was associated with a 4.6 percentage points lower probability of graduating high school on time. Skipping at least one question in each of the exams was associated with an almost 6 percentage points lower probability of graduating on time.
Soland (2018)	85 schools taking the OECD Test for	Item Durations	Students with high self-efficacy spent 1.5 times as long on difficult items and were 18 times less likely to rapidly guess on those items	
Soland, Jensen, Keys, Wolk, & Bi (2018)	Measures of Academic Progress (MA	Rapid Guesses	Partial correlations with self-management of .26 and with self-efficacy of .12.	On average, students disengaged on the test were absent 1.3 more days per year, 3 times as likely to have a detention, 4 times as likely to have a suspension, and had GPAs that were .8 points lower
Swerdzewski, Harmes, & Finney (2009)		Rapid Guesses	Rapid guessing associated with feelings of academic autonomy, feelings of academic competence, interest in academics, and enjoyment in academics	

Table 3

Summary of Findings on The Relationships among Survey Metadata, SEL, and Academic Engagement

Study	Data	Metadata Source	Findings		
			SEL	Academic Engagement	Later Life Outcomes
Cheng (2015)	Longitudinal Study of American Youth: 1987	Item Nonresponse			A standard deviation increase in item nonresponse rate in middle school is associated with completing a 0.5 fewer years of education and a 4 percentage point increase in the likelihood of being employed at age 36, net of cognitive ability and demograhpic background charateristics.
Hitt (2016)	National Educational Longitudinal Study: 1988 Educational Longitudinal Study: 2002	Careless Answers Careless Answers	Raw correlations of -0.24 and -0.10 with locus of control and self-efficacy, respectively.		A standard deviation increase in careless answering in middle and high school is associated with completing 0.8 fewer years of education by age 26.
Hitt, Trivitt & Cheng (2016)	High School and Beyond: 1980, National Longitudinal Study of Adolescent to Adult Health (Add Health), National Longitudinal Study of Youth: 1997, Educational Longitudinal Study: 2002	Item Nonresponse			A standard deviation increase in item nonresponse rate in middle and high school is associated with completing a 0.1 to 0.3 fewer years of education by about age 26, net of cognitive ability and demograhpic background charateristics.
Zamarro, Cheng, Shakeel, & Hitt (2018)	Understanding America Survey	Item Nonresponse; careless answering	Partial correlations of careless answering with conscientiousness and grit are about -0.15. Partial correlations of item nonresponse with conscientiousness and grit are about 0.05.		
Zamarro, Nichols, Duckworth, & D'Mello (2017)	Longitudinal data from a convenience san	Item Nonresponse; careless answering	Partial correlations with self-management and grit ranging from -0.2 to -0.17.	A standard deviation increase in item non-response and careless answering are associated with a 0.2 and 0.17 standard deviations decrease in senior year GPA and are 23 and 13 percentage points less likely to attempt the SAT, respectively	A standard deviation increase in item non-response and careless answering are associated with 24 and 10 percentage points lower probability of enrolling in college for freshmen year, respectively