Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

EVALUATING BIAS: CONSIDERING ALTERNATIVE MEASURES OF GLOBAL CITIZENSHIP

Michael Thier, Ross Anderson, Christine Pitts
Department of Educational Policy, Methodology, & Leadership
University of Oregon, Eugene, OR

Abstract

Responding to a groundswell of researcher, policymaker, and practitioner interest in developing students' metacognitive skills, we evaluated measurement approaches for global citizenship. We designed a 10-criteria evaluative framework from seminal and cutting-edge research to compare an extant self-report, the Global Citizenship Scale, and situational judgment test, the Virtual Cultural Awareness Trainer. We also designed a discrete choice experiment for global citizenship. Our evaluation illustrated strengths, limitations, and trade-offs presented by each approach's design considerations, possibilities for bias, and validity issues. We found that researchers rely almost exclusively upon self-report instruments to measure global citizenship and many other metacognitive skills despite those issues. Self-report instruments are susceptible to biases, which partially explain why psychometricians have privileged measuring cognitive over metacognitive domains. This study shows all three measures to mitigate response-style bias and to lack sufficient data for internal and criterion-related validity. On the whole, the SJT and DCE offered slightly more strengths than the self-report, but we deem none ready for use independently within K-12 accountability systems. We call for further development of all three measurement alternatives.

Keywords*: discrete choice experiments, measurement bias, metacognitive skills, self-reports, situational judgment tests*

Introduction

Researchers and educators refer to groups of skills as *social/emotional* (Waters & Sroufe, 1983), *non-cognitive* (Heckman, 2000), *metacognitive* (Vockell, 2009), and/or *21st-century* (Soland, Hamilton, & Stecher, 2013). We use "metacognitive" because the term facilitates complex understanding of skills that require high-level processing, strategy selection, and reflective thinking (Conley, Beach, Thier, Lench, & Chadwick, 2014). By any name, metacognitive skills—such as those we exemplify in Table 1—span academic disciplines, encompass lifelong dispositions and aptitudes, and foster "a learners' automatic awareness of their own knowledge and their ability to understand, control, and manipulate their own cognitive processes" (Vockell, 2009). Scholars have attributed to metacognitive skills (a) greater utility for 21st-century students than discrete cognitive skills such as mathematics and reading (National Research Council, 2012); (b) being at least equal to cognitive skills in their ability to predict postsecondary success (Conley, 2013); (c) greater malleability than cognitive skills (Heckman, 2000), and

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

(d) better ability to predict long-term academic and economic outcomes (Landine & Stewart, 1998; Lindqvist & Vestman, 2011; Vrugt & Oort 2008).

Despite these benefits, many developed nations have eschewed their students' metacognitive skills in favor of chasing standardized exams scores (Zhao, 2012). Recently though, educators worldwide have begun a groundswell toward developing metacognitive skills (Farrington et al., 2012; Zhao, in press), perhaps because they require deeper learning than rote tasks (Conley et al., 2014). Multinational businesses clamor for applicants who are creative, collaborative global citizens and approach tasks with grit and a growth mindset (Farrington et al., 2012; Soland et al., 2013). Regardless of the high value that employers and societies place on these skills, few schools measure them (Rothstein, 2004). If schools do not measure metacognitive skills, we can assume few schools prioritize their development even though metacognitive assessments can provide actionable information in ways that more common cognitive assessments cannot (Conley, 2013).

Table 1

Sample of Metacognitive Skills Defined

Skill	Definition	Citation
Creativity	Originality; something novel, unique, unusual, and distinct from what we expect to experience or have experienced; effectiveness, usefulness, fit, appropriateness, and value of the creative act, product, or idea	Runco & Jaeger (2012)
Collaboration	Communication, plus additional competencies related to conflict resolution, decision making, problem solving, and negotiation	Lai (2011)
Global citizenship	Social responsibility, global competence, and global civic engagement	Morais & Ogden (2011)
Grit	Perseverance; passion for long-term goals; and persistence through challenge	Duckworth, Peterson, & Matthews (2007)
Growth mindset	Seeing one's intelligence as malleable and a function of effort; contrasts with a fixed mindset that treats intelligence as an innate ability, immune to the efforts of an individual to improve	Dweck (2010)

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

However, schools might be justified for their reluctance to embrace teaching and assessing metacognitive skills. Their diffuse and overlapping definitions problematize their measurement (Soland et al., 2013). So policymakers should not rashly create the conditions to entrench metacognitive skills in the landscape of K-12 outcomes without first learning their measurement challenges (Duckworth & Yeager, 2015). Principally, attempts at metacognitive measurement have exposed biases inherent in self-report instruments (Roberts, Martin, & Olaru, 2015). Such biases include social desirability, egocentrism, and stereotype threat (Ahmed, Van der Werf, Minnaert, & Kuyper, 2010; Kopcha & Sullivan, 2007; Lagattuta, Syfan, & Bamford, 2012), which thwart attempts at accurate measurement. For example, a student who can intuit the ideal attributes of a metacognitive construct might fake those qualities on an assessment (Huws, Reddy, & Talcott, 2009). Problematically, increasing stakes attached to an assessment will increase the likelihood of faking (Conley et al., 2014; Duckworth & Yeager, 2015). Such potential for bias explains partially why psychometricians have traditionally privileged measuring cognitive over metacognitive domains (Conley, 2013). As a result, educators have missed opportunities to change practices and affect student outcomes. Metacognitive assessments identify achievement gaps as functions of student effort rather than aptitude, so such diagnostic ability can be particularly useful for practitioners of responsive pedagogy (Pecheone, Kahl, Hamma, & Jaquith, 2010).

One obstacle that metacognitive advocates do not suffer from is a lack of extant measures. Soland et al. (2013) called the number of options "dizzying" (p. 9), but they noted that the types of skills defined in Table 1 do not align neatly with traditional assessment such as standardized tests of basic literacy and numeracy. Still, cognitive skills have received far more attention from psychometricians due in part to resources from a robust testing industry, causing a misalignment between domains that might be clearer for measurement but less useful for educators (National Research Council, 2012). Conley et al. (2014) noted another factor: standardized tests. Debuting in the 1950s, standardized tests have woven themselves into the fabric of school in many developed nations. As a result, metacognitive measures are:

> not seen as technically rigorous, are not used, and therefore the technical rigor is never improved. Use remains limited to boutique schools and esoteric settings, and the general

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

> public's familiarity with these instruments never increases. (p. 7)

However, educational systems in about 20 American states have begun to implement metacognitive assessments (Conley, 2015; Evans-Brown, 2015; Partnership for 21st-Century Skills, n.d.). Given that forward-thinking practitioners and policymakers need additional metacognitive measurement options, the current study compared three measurement approaches to global citizenship to illustrate strengths, limitations, and trade-offs. One measurement approach, self-report, has a century of experience both within education and in other fields of social science research. The others, situational judgment tests and discrete choice experiments, have established histories in other fields, but are first emerging within education research.

We selected global citizenship because it has been identified as a goal of both K-12 (Zhao, 2010) and higher education (Morais & Ogden, 2011), but it has been operationalized and measured much more frequently in higher education (Thier, in press). The combination of unequal access to higher education and the rarity of K-12 programs that prioritize global outcomes (e.g., International Baccalaureate) suggest an opportunity gap for economically disadvantaged students (Killick, 2012; Reimers, 2010; Thier, 2015). Still, the prevalence of globalization creates a need for all K-12 public schools to provide global education (Duncan, 2013; Molina & Lattimer, 2013) and measure progress toward global citizenship.

Definitions

We developed a framework to evaluate a *self-report* and *situational judgment test* (SJT) of global citizenship, and guide our development of a *discrete choice experiment* (DCE). Of the three measurement approaches, self-report is the most commonly used to collect data in social science research (Weijters, Geuens, & Schillewaert, 2010). Typical examples of self-report measures include surveys or questionnaires in which respondents answer questions independent of researchers. SJTs measure procedural knowledge within specific domains without completing formal, field observations of respondents (Lievens & Sackett, 2012). SJTs include hypothetical scenarios from which respondents rank

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

or identify the most appropriate responses based on their feelings and/or knowledge. Researchers solicit respondent reactions and estimate the choices respondents would likely make in reality. DCEs present respondents with hypothetical choices from varying attributes within choice sets. Respondents state their preferences. Ultimately, researchers estimate the value of each choice and the attributes of those preferences (Kennelly, Flannery, Considine, Doherty, & Hynes, 2014).

Methods

To establish the initial literature pool for this review, we searched ERIC for peer-reviewed articles from 2006-2015 with combinations of the following keywords: self-report, situational judgment test, discrete choice experiment, and bias. We found 206 articles: 95 results for self-report bias, 81 for SJT bias, and 30 for DCE bias. We included only articles that had titles, keywords, or descriptors with measurement, validity, and/or reliability, yielding 15 studies. We expanded our pool by adding a seminal study by an author that all of our other DCE articles under consideration had cited in their methods' sections. We synthesized the 16 studies to develop a framework for evaluating strengths, limitations, and trade-offs of the three measurement approaches. Our review informed construction of an evaluative framework with three dimensions: (a) measurement design, (b) possibilities for bias, and (c) validity. In Table 2, we defined each dimension. Using this framework, we evaluated global citizenship measures based on potential for use with high school students.

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

Table 2

Framework for Evaluating Approaches to Metacognitive Measurement

Dimension	Subdimension	Definition
Measurement design		The decisions each measure's creator(s) made, reported making, or failed to make during development
	Preparatory qualitative measures	Qualitative research used during the design process to ensure a larger and more contextually appropriate range of a construct, to match the respondent population, and to increase the validity of findings
	Measures of internal consistency	Traditional criteria (e.g., Cronbach's alpha), as well as surveys, choice experiments, interviews, and focus groups used to increase reliability
	Efficiency trade-offs	Decisions designers use to weigh modeling considerations for statistical efficiency and respondents' increased cognitive load
Possibilities for response bias		Common factors that influence responses to items
	Self-presentation	When measured results are inaccurate due to respondents' tendencies toward socially desirable responses that show them as greater than actual
	Egocentric	Respondents use their own varying viewpoints to determine response choices, bringing their own context to the measure (related issues include: experiential bias, reference bias, and intrapersonal harshness)
	Stereotype threat	Respondents experiencing the risk of confirming, as self-characteristic, a negative stereotype about one's group, biasing findings as a respondents' emotions affect their judgments and appraisals
	Response style	Systematic inconsistencies within responses over items in a single measure
Validity		Characteristics that enhance or decrease a scale's ability to measure what it intends to measure
	Internal	Extent to which a conclusion from a study is warranted based on how a study minimizes systematic error
	External	Extent to which one can generalize conclusions from the findings
	Criterion-related	Measure's ability to predict or measure concurrent performance

Measurement design

In this dimension, we focused on the decisions each measure's designers made, reported making, or neglected to make during development. We examined *preparatory qualitative measures*, *measures of internal consistency*, and *efficiency trade-offs*, all of which signal decisions about model specifications, item construction, reliability, and design.

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

Preparatory qualitative measures

Qualitative measures can be used during the design process to ensure inclusion of a more encompassing range of interests, to match respondent populations, and to increase the validity of inferences from findings. A designer's assumptions might limit the available options or factors included in the information solicited (Cunningham et al., 2009). Focus groups, or other qualitative approaches, can help designers specify characteristics to include during item-construction processes (Kennelly et al., 2014) as can wording of items or number and labeling of levels on scales (Cunningham et al., 2009).

Measures of internal consistency

Traditionally, designers rely on correlations of items to determine internal consistency, most frequently Cronbach's alpha (Osterlind, 2009). However, some researchers use surveys, choice experiments, interviews, and focus groups to decrease their rates of false positives (Taylor, Vehorn, Noble, Weitlauf, & Warren, 2014). When multiple measures are not feasible, Taylor et al. employ internal metrics of response characteristics to examine bias during test construction. Similarly, Kennelly et al. (2014) advocate status quo options within DCEs to avoid trade-offs during situations in which respondents might not be familiar with a context, and therefore, unwilling to make forced choices.

Efficiency trade-offs

Employers of measures that require respondents to make choices, comparisons, or judgments must consider trade-offs between modeling statistical efficiency and respondents' cognitive load. Often, SJTs measure novel tasks that might be especially demanding if a respondent has not experienced them before. A forced choice approach adapts the DCE method, limiting the number of attributes considered in a single choice set, though requiring more sets for reliability (Roberts et al., 2015). Cognitive demand becomes increasingly burdensome as task complexity increases in either number or kind of characteristics. For DCEs, computer software can define attributes and levels of choice, thus controlling for respondents' cognitive load (Kennelly et al., 2014). In fact, Louviere, Islam, Wasi, Street, and Burgess (2008) caution that respondents typically choose the most cognitively effective strategy to answer

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

questions, which results in error variance that could bias estimates, a facet designers should consider when examining response patterns.

Possibilities for response bias

In this dimension, we focused on factors that might bias results such as: *self-presentation, egocentric*, *stereotype threat*, and *response style*.

Self-presentation

This type of bias occurs when measures produce inaccuracies due to respondents' tendencies toward that which is socially desirable, revealing the respondent to be "greater-than-actual" (Kopcha & Sullivan, 2007, p. 14). The result is *faking* (Huws et al., 2009; Roberts et al., 2015). Self-reports, which are commonly used due to convenience and low cost, are particularly susceptible to this type of bias (Duckworth & Yeager, 2015). Self-presentation has been measured on socially sensitive topics only, but Miller (2012) urges designers to explore this bias when developing any measure.

Egocentric

Instances in which respondents use their varying viewpoints to determine response choices, ostensibly bringing their own context into measures, could bias responses (Cunningham et al., 2008; Lagattuta et al., 2012). For example, Lagattuta et al. measured the convergence of child- and parent-reported emotions, finding that increases in parents' ratings of their own worries/optimism increased *their* ratings of *their child's* worries/optimism. Similarly, the choice options presented in DCEs might reflect only a small portion of the real-world decisions that respondents make (Cunningham et al., 2008). As a result, respondents' varying contextual backgrounds might create biases for which researchers' modeling procedures do not always account (Waschbusch et al., 2011). Precise sampling procedures and preparatory qualitative measures can reduce susceptibility to this bias.

Stereotype threat

The extent to which respondents experience stereotype threat—the risk of confirming, as self-characteristic, a negative stereotype about one's group (Steele & Aronson, 1995)—can bias findings

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

because respondents' emotions can affect their judgments and appraisals. Judgments about performance on new tasks are more susceptible to stereotype threats than tasks situated in one's typical conditions (Howard & Anderson, 2010). Examining emotions and appraisals of a math lesson, Ahmed et al. (2010) found students to experience emotions and associate them directly with their self-reports of competence and the value they found in lessons, positively and negatively. Using SJTs, however, Howard and Anderson (2010) examined students performing novel tasks while their race/ethnicity was stigmatized. The researchers found low respondent-expected performance, but actual performance did not decrease.

Response style

A systematic assessment of inconsistencies within responses over items in a single measure might yield response-style bias (Weijters et al., 2010). Most designers consider inconsistencies in responses to be unsystematic and attribute them to random error. However, this bias might occur if a measurement disproportionately uses positive responses (i.e., acquiescence response style) and/or extreme responses. Such response patterns should not be attributed to random error. Weijters et al. suggest diagnosing and correcting such biases by capturing variance through the use of reverse-coded items and other methods.

Validity

In this dimension, we focused on characteristics that enhance or decrease a scale's ability to measure what it intends to measure. We emphasized *internal*, *external*, and *criterion-related* validity.

Internal and external

Shadish, Cook, and Campbell (2002) emphasize internal validity, the extent to which a conclusion from a study is warranted based on how a study minimizes systematic error, and external validity, the extent to which one can generalize conclusions from the findings. Designers must consider that both depend upon several factors, including research design and sampling strategies. Even though DCEs enable modeling of respondents' choice trade-offs, thus reducing social desirability bias and better estimating actual behaviors (Cunningham et al., 2009; Kennelly et al., 2014), poor measurement design might thwart generalizability. Concepts such as cost or value might be context-bound, resulting in

different inferences across respondents (Kennelly et al., 2014). Also, the specificity of the population used to sample can limit generalizability (Waschbusch et al., 2011). All three measurement approaches under consideration in this study reflect single time points of respondent perspective or behavior. No single measure can capture the dynamism of evolving attitudes, characteristics, and/or contexts (Cunningham et al., 2009).

Criterion-related

When measures forecast future performance in other criteria (e.g., selection decisions) or correlate with concurrent measures, they achieve criterion-related validity (Thorndike & Thorndike-Christ, 2010). Self-report surveys of preference and SJTs are frequently used to make decisions (i.e., job or school placement). Lievens and Sackett (2012) found SJTs to predict job and internship performance better than cognitive factors, or self-report alone, supporting SJTs' potential for criterion-related validity. However, measuring novel tasks creates an inherent threat to the predictive nature of SJTs due to their greater susceptibility for egocentric and stereotype threat biases (Howard & Anderson, 2010).

Results

We applied the evaluative framework to reliability and validation studies of an extant self-report and an SJT that can be used to measure global citizenship, as well as our design of a DCE. Morais and Ogden (2011) designed their 30-item self-report, the Global Citizenship Scale (GCS) with a 5-point, Likert-type scale to measure higher education study abroad outcomes. The GCS is among 140-plus extant measures of constructs related to global citizenship—nearly all self-reports (Deardorff, 2014). As SJTs are less common approaches to measurement, we applied the framework to the Virtual Cultural Awareness Trainer (VCAT), a computer-based simulation in which users employ avatars to develop proficiency in intercultural communication, problem solving, and cultural knowledge (Johnson, Friedland, Schrider, Valente, & Sheridan, 2011). Last, DCEs just entered educational research (see Aubusson, Burke, Schuck, Kearney, & Frischknecht, 2014) so no studies have used DCEs yet to measure global citizenship. Rather, we examined the strengths, limitations, and trade-offs of the GCS and VCAT (see

Table 3) to design a DCE for measuring global citizenship. DCEs have been used across various sectors to understand the power of individual preferences in predicting future consumer behavior. Regarding students as the ultimate consumers of educational opportunities, we posit that DCEs can estimate students' metacognitive skills.

Table 3

Strengths, Limitations, and Trade-offs of Self-Report and Situational Judgment Test of Global Citizenship

Measure	Strengths	Limitations	Tradeoffs
Global Citizenship Scale (Morais & Ogden, 2011)	Thorough audit trail of validity/reliability processes; two highly reliable dimensions; attempts to reduce response style bias	Social responsibility as "unclear dimension;" Westernized egocentric bias; potential for selection and history threats to validity	Recent attempts to expand external validity v. dilution of expert-established construct validity
Virtual Cultural Awareness Trainer (Johnson et al., 2011)	Potential for respondent to self-moderate cognitive load and generalizability; avoids stereotype threat and response style bias	No peer-reviewed studies; infeasibility of conducting traditional validity/reliability studies; susceptibility to egocenteric and self-presentation biases	Unorthodox approach to assessment; tension between formative and summative uses; respondent engagement vs. validity

Self-report

GCS is the first scale to operationalize global citizenship by name, defining it with three overlapping dimensions: social responsibility, global competence, and global civic engagement.

Measurement design

Morais and Ogden (2011) provided transparent reporting of the methods they used to design the GCS, supporting their claims of theoretical grounding and empirical validation. Their eight-step process included multiple expert face-validity trials, exploratory and confirmatory factor analyses, and nominal group technique interviews. The authors do not report how they culled the citations in their literature review, but their pool spanned nearly 20 years of peer-reviewed publications. They incorporate 12 extant measures that partially operationalized at least one global citizenship dimension. Morais and Ogden employed *preparatory qualitative measures*, seeking discrepant views (see Merriam, 1998) by convening two expert panels for item review. They invited subject experts ($n \sim 80$) to categorize initial items by

hypothesized dimensions, provide item-level feedback, and exclude any items not assigned consistently to a dimension. The process yielded 43 items, 13 each for social responsibility and global competence, and 17 for global civic engagement. They streamlined the scale to 30 items through factor analyses and solicited feedback from 25 American university students who had studied abroad.

Factor analyses revealed strong *measures of internal consistency* for the GCS. Its final, higher-order, 10-factor model met standard model-fit criteria (see Osterlind, 2009). The authors reported that effect sizes for each parameter surpassed .10 ($p < .01$), but did not report effect sizes individually. To assess reliability, they conducted a split-half test. Retained items produced a strong coefficient (.91). Morais and Ogden reported Cronbach's alphas across GCS's six subscales ranging from .67-.92. Paths from global competence (.77) and global civic engagement (.99) to global citizenship were strong. Generally, Morais and Ogden retained items with factor loadings above .50, but they did not provide a clear decision rule for the 12 instances in which they retained or excluded items that did not conform to that specification. They did, however, retain one social responsibility item with a loading of .21 because it aligned "theoretically" with Factor 1 (p. 456), the only factor that loaded onto social responsibility, albeit weakly. Social responsibility had a weak path to global citizenship (.13). Morais and Ogden reverse-coded all 6 social responsibility items, forward-coding the 24 items in the other two dimensions. The authors called social responsibility "an unclear dimension" (p. 461). Based on factor analysis and interviews, social responsibility was the only GCS dimension for which a hypothesized multifactor structure did not hold.

Possibilities for response bias

Compared to their thorough reporting of development procedures, Morais and Ogden did not offer as comprehensive a discussion of GCS' potential for bias. Morais and Ogden assert the GCS is less susceptible to the type of *self-presentation bias* found in other self-report measures of study abroad outcomes because it relies on self-assessment in the present, not retrospective experiences. In fact, 12 items are future-oriented (e.g., I plan to get involved in a program that addresses the global environmental crisis). Still, the authors validated their scale in part by sampling students in classrooms at five Penn State

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

University campuses. Lombardi, Seburn, and Conley (2011) found student survey-takers in classrooms to be susceptible to perceiving teacher incentives for favorable responses even if they do not reflect reality, especially when the survey seeks to capture student beliefs about their own knowledge, skills, and attitudes. Such an effect might be even more salient in K-12 than among university students.

Morais and Ogden noted a limitation of the GCS: the literature that informed its design came disproportionately from scholars in the Western hemisphere and relied exclusively on data from North American practitioners, academics, and students. The authors called for further research to broaden the GCS' scope and potential uses. In the interim, respondents whose viewpoints do not align to Western norms might bias results because their reference points do not necessarily conform to the scale's operationalization. To that point, Lawthong (2003) administered the Hett (1993) Global Mindedness Scale—one of the 12 extant measures Morais and Ogden used to form their item pool—to 1,739 public university students in Thailand. When using Hett's U.S.-developed global-mindedness scale, Lawthong found the Thai socio-cultural context to account for 86% of the variance. Morais and Ogden acknowledge that the GCS might reflect a Western slant. As such, it could create the conditions for *egocentric bias* and *stereotype threat*, particularly among K-12 students who do not identify with the type of cultural norms often codified in American schools. Finally, Morais and Ogden edited items that were lengthy, double-barreled, or contained ambiguous pronoun references, all efforts to reduce *response-style bias*.

Validity

Morais and Ogden stated that modification would make the GCS ideal for pre-/post-test designs. Such usage should prompt researchers to consider selection and history, two threats to *internal validity* (Shadish et al., 2002) as they design and/or model statistics with the GCS. When the authors established reliability by sampling college students, their data depended upon individuals whose selection into the study came at a time of likely identity formation and/or reformation. As such, their scores on a self-report, attitudinal scale carry potential for unmeasured confounds. For example, one's global citizenship orientation might increase or decrease on the basis of age, creating a plausible history threat for pre-/post-test designs, especially given long intervals between measurement occasions. Therefore, researchers using

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

GCS thusly might mistake treatment effects (e.g., from study abroad) for artifacts of concurrent events (e.g., natural maturation). Importantly, Morais and Ogden consider their measure sufficient for research and practice with study abroad programs. They sampled both students who resided overseas during their programs and who had short international visits within domestically embedded experiences, allowing for greater *external validity* across American college campuses. Ongoing studies might facilitate even wider reach for the GCS. One study seeks to adapt the scale for Turkish respondents; two others sampled business leaders in Maine and elementary school educators in South Carolina (personal communication, F. Cermik; K. Tardiff; S. Durham). These studies might provide pathways to use the GCS in a larger number of settings.

Situational Judgment Test

VCAT's designers do not use the terms SJT or global citizenship explicitly, but their interactive role-playing scenarios train users in intercultural interactions "to make decisions about how to proceed in response to different types of unfamiliar, desirable, or undesirable reactions from non-player characters" (p. 5-6). The VCAT is consistent with multiple-choice SJTs (Roberts et al., 2015). No peer-reviewed journal has published information on the VCAT, one of the computer simulations made by Alelo, Inc., which began its research at the University of Southern California. Instead, we relied on conference and white papers, plus internal reports. In a white paper from RAND, Soland et al. (2013) endorse Alelo products as metacognitive assessment options, using the caveat that

> the very definition of assessment is relaxed. Students do not take a formal test of any kind; rather, they receive constant feedback on their performance, which they can then use to improve at their own rate" (p. 29)

By design, Alelo conceived most of its products to train military personnel for deployment into unfamiliar linguistic and cultural settings, not K-12 use. Taking the progressive view that VCAT can test metacognitive skills in K-12 schools, Soland et al. emphasize that "traditional reliability and validity estimates are not especially feasible" (p. 43) to date for Alelo's products.

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

Measurement design

Johnson et al. (2011) used *preparatory qualitative measures* during development, combining ethnographic interviews and other unspecified "anthropological and linguistic research" (p. 3) and "best-practice data collection methods" (p. 4). By contrast, little publically available data provides information on Alelo's *measures of internal consistency*. Hanson (2013) reported a cost-cutting study of the Danish Simulator, which Alelo based on its Tactical Language and Culture Training Systems. Hanson does not offer sample sizes or procedures, but reports having tested two groups of language learners, one with a classroom teacher and one with the Simulator. In another study providing limited data, Johnson and Zaker (2012) described the virtual coach, who guides users through Alelo simulations, providing feedback when users make linguistic/cultural errors or social faux pas. The guide explains how to avoid future problems, a feature that enhances learning, but threatens test-retest reliability. Though not explicitly stated, Alelo seems to have thought considerably about *efficiency trade-off*. Its instructional model consciously balances intent to provide real-time feedback with a desire to reduce cognitive load. When users err, simulated native speakers alert users overtly or through body language. Soland et al. (2013) recognize that such subtleties enhance the "real-world scenarios" allowing for real-time student assessment, "albeit more overtly than might occur conversationally" (p. 43) in the physical world. This approach reduces cognitive load, consistent with Alelo's intentional blend of curriculum and measurement.

Possibilities for response bias

Alelo products seem susceptible to *self-presentation* and *egocentric* biases but seem designed to minimize *stereotype threats* and *response style* biases. Problematically for measurement, VCAT users complete questionnaires that determine which modules they receive. Astute users might detect socially desirable questionnaire responses, thus rigging their VCAT experiences toward self-presentation. Similarly, users can disable real-time feedback, opting for summative virtual coaching instead. Such a choice might eliminate important data about judgments, the strength of SJTs (see Lagattuta et al., 2012). Positively, VCAT's whole reason for being seems to be for teaching users to navigate stereotype threats that exist in the physical world. For example, Alelo's simulators signal users when characters' attitudes

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

change in intercultural interactions. Such changes influence learner responses (Johnson & Wu, 2008, p. 3), creating opportunities to recognize, understand, and react to potential stereotype threats such as practicing different and appropriate ways to greet an old man, a peer, and a little girl (Johnson & Zaker, 2012). Furthermore, Alelo designed the signals and virtual coaches' interventions to account for the "affective state of the learner" and avoid discouraging "face-threatening feedback" (Johnson et al., 2011, p. 5). It is possible that VCAT learners experience less stereotype threat than some students in traditional classrooms. Unlike an online space, the diverse contexts of schools in many developed nations show considerable evidence that reifying positive and negative stereotypes affects student performance (Aronson & Dee, 2012). Furthermore, Alelo addresses the possibility of response style bias because it does not follow rigid curriculum. Instead, Alelo offers a variability function that facilitates users to develop the disparate cultural skills one might need to patrol a checkpoint on the Afghan border or coordinate a humanitarian aid project with host-nation officials (Johnson et al., 2011).

Validity

With respect to *internal validity,* Soland et al. (2013) considers Alelo products as multiple measures of "overlapping competencies, such as critical thinking, academic mastery, and communication" (p. 29). Though highly useful for instructional purposes, such overlap presents potentially insurmountable measurement challenges. How researchers should parse constructs and meaningful variance, wraps VCAT in a black box. By contrast, VCAT shows promise for *external validity*. Though information on VCAT's K-12 applicability is limited (Soland et al., 2013), Alelo's simulations can facilitate transdisciplinary lessons that would be ideal for 21st-century learners. Regarding *criterion-related validity*, VCAT's evidence is mixed. Alelo opposes "culture-specific training" (Johnson et al., 2011, p. 3). Instead, Alelo develops culture-general skills and attitudes; VCAT's designers tag the separation of global competencies into culture-specific bins as artificial. Alelo teaches perspective-taking and rapport-building generally, because such skills apply in "a wide range of contexts" and demand simulations to be "concrete and understandable" (p. 3). However, Alelo's conceptualization counters notions that citizenship learning, especially in a global sense, is "complex and occurs in a wide variety of formal, non-formal and informal

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

learning settings" (Eidoo et al. 2011, p. 59). Still, Johnson et al. (2011) consider their purpose to "develop attitudes necessary to apply [users'] cultural knowledge in situations that they are likely to encounter" (p. 2). Alelo sampled active-duty and former military personnel, domain and occupational experts, and native citizens and/or speakers from target geographical areas. Still unknown: Does such sampling enhance concurrent and predictive validity or only face validity?

Discrete choice experiment

We designed a DCE (see Figure 1) to offer an alternative to self-report and SJT approaches. Our design stemmed from a literature-based evaluative framework. We envision use of our DCE such that respondents choose from attribute combinations that researchers vary experimentally (Cunningham et al., 2009). Ultimately, each attribute would receive a weighted value (i.e., a regression coefficient) for each choice outcome. In the ensuing section, we consider the measurement design of our DCE and the resulting considerations for bias and validity.

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

Figure 1: Discrete Choice Experiment to Measure Social Responsibility

Scenario: Your class received two exchange students from different countries. Both exchange students requested you as a partner on a project your teacher just assigned, one that will not be graded. It is just for learning's sake. Your teacher assigned the exchange students to think up the project. Then, your teacher asked you to choose which project/partner to join. Read the two options about your potential partners and projects to determine which you will join. Please answer the follow-up question at the bottom of the page.

	Attribute characteristics for choice sets	
Attributes	Low socially responsible alternative	High socially responsible alternative
Sense of global justice	I want to work on a project with easier work or an easy-to-work-with partner.	Even if a project is hard, or a partner is hard to work with, I want to contribute to goodness.
Recognition of disparities	I want to work with a partner who has a reputation for being successful.	Despite reputation, I want to work with a partner whose life has not been easy.
Altruism	I want to work with a partner who completes a fair share of a project.	I want to work with a partner who might need additional help to complete a project.
Empathy	I want to work on a project if it serves my interests.	I want to work on a project that teaches me about someone else.
Awareness of interconnectedness	I want to work on a project that addresses concerns near home.	I want to work on a project that addresses concerns near home and in places I've never been.
Feeling of personal responsibility	I want to work on a project that seems like it will be fun.	I want to work on a project that has potential to help people.

Please rate overall how well the option you selected fits your style from 1 (not a good fit) to 5 (very good fit).	1	2	3	4	5

To explore the DCE as a measure of global citizenship, we referred to the GCS (Morais & Ogden, 2011) and its theoretical foundations. We aimed to measure social responsibility as 6 attributes: sense of global justice, recognition of disparities, altruism, empathy, awareness of interconnectedness, feeling personal responsibility based on the GCS items in that domain. We selected social responsibility because the authors reported it as an unclear dimension. We hypothesized that a DCE might present a useful alternative for measuring social responsibility. We note our DCE's potential confound between social responsibility and collaboration, though Soland et al. (2013) regarded metacognitive skills as inherently overlapping, even across domains. Furthermore, scholars of global competence—a construct Singh and

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

Qi (2013) report as interchangeable with global citizenship and 10 other terms—consider collaborating on multinational/multicultural teams to be an essential component (Brustein, 2009; Hunter, 2004; Parkinson, 2009; Willard, 2009). Our DCE uses a secondary question to analyze the weighted contributions of each attribute to the "best fit" choice outcome. This final question also requires that respondents reconsider their choices and re-examine the nature of each attribute across options in the choice set. In similar fashion to reverse-coded items items, this *measure of internal consistency* serves the dual purpose of analytical importance and reduction of *response style biases*. Responders must weigh options carefully. To balance cognitive load with capturing social responsibility fully, we capped levels per attribute at 2. Past research on DCEs found statistical efficiency (e.g., more items but smaller sample) might sacrifice respondents' choice consistency, leading to higher error variance (Louviere et al., 2008). Assumptions inherent in each design decision require testing for statistical feasibility alongside issues of cognitive load.

For reducing bias, many benefits accompany DCEs. First, by controlling variation in one attribute of the construct, we can ensure it is not correlated with another. Second, DCE's hypothetical scenarios can parse observable from unobservable data. One's preferences or perceptions tend to correlate strongly with observable data. Third, by choosing scenarios carefully and describing attributes realistically, we can elicit choice that draws connections for respondents, enabling us to model perceptions and preferences more comprehensively than correlational measures (see Aubusson et al., 2014). By forcing choices upon respondents, we can expect DCEs to reduce response biases, revealing more realistic perceptions or preferences when compared to self-report. Respondents compare attribute differences based on detailed descriptions rather than ambiguously termed ratings, common to Likert-type scales (e.g. "not at all," "somewhat," or "very much"). Therefore, we can expect DCE respondents' *self-presentation* or *egocentric biases* to decrease. Ultimately, harnessing DCE's potential requires additional investigation before they could be recognized as allowing for valid inferences about students' metacognitive skills.

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

Discussion

In this study, we built an evaluative framework to compare three approaches to measuring metacognitive skills, using global citizenship measures to exemplify. Our literature search demonstrated the saturation of self-report as the measure of choice across multiple constructs. Far more empirical investigation has examined strengths and limitations of self-report, both in the social sciences generally and for global citizenship and related metacognitive construct specifically (see Fantini, 2009). After synthesizing literature, we categorized issues to consider when approaching measurement of metacognitive skill. We varied the framework slightly across self-report measures, SJTs, and DCEs to account for the emergent status of the latter two approaches, particularly within research on K-12 education and metacognitive skills. In Table 4, we summarized preliminary evidence for the promise of SJTs and DCEs.

Table 4

Strengths, Limitations, and Potential for Approaches to Measuring Global Citizenship

Criterion	Global Citizenship Scale	Virtual Cultural Awareness Trainer	Discrete Choice Experiment
Preparatory qualitative measures	Strength	Strength	Strength
Measures of internal consistency	Strength for 2 of 3 dimensions	Limitation	Potential
Efficiency trade-offs	No data	Strength	Strength
Potential for self-presentation bias	Limitation	Limitation	Strength
Potential for egocentric bias	Limitation	Limitation	Strength
Potential for stereotype threat	Limitation	Strength	No data
Potential for response style bias	Strength	Strength	Strength
Internal validity	Limitation	Limitation	No data
External validity	Strength	Strength	No data
Criterion-related validity	No data	Potential	No data

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

No approach provides a silver bullet for measuring global citizenship. All three measures share strengths in their preparatory qualitative measures and mitigation of response-style bias. All three present limitations or lack sufficient data to endorse their internal and criterion-related validity. For researchers, policymakers, or practitioners who want a measure that offers the best internal consistency at present, the GCS is the best option. If a user seeks to manage the potential for stereotype threat, the VCAT would be preferable. For users who want to minimize potential for self-presentation or egocentric biases, our DCE seems to be the strongest. The GCS and the VCAT share strong external validity that our DCE does not. The VCAT and our DCE seem to have addressed efficiency trade-offs; the GCS has not. Nominally, however, the VCAT and DCE offered an additional strength when compared to the self-report. Though, none of these measures is ready to be used on its own for accountability or other high-stakes decisions. Our results confirm scholarship about self-reports across metacognitive domains (see Conley et al., 2014 and Duckworth & Yeager, 2015); they are not measurement panaceas.

The brimming question—*How can we measure a metacognitive skill such as global citizenship?*–might require a multi-method approach that includes all three measures. Also, featuring multiple reporters can further detect, reduce, or control for biases. Furthermore, complex constructs such as global citizenship contain value- and culture-laden norms and expectations (Morais & Ogden, 2011). Applying our evaluative framework to other measures of global citizenship and to other metacognitive constructs is a necessary next step. For example, in a companion paper to this study, we used the same framework to examine a self-report and SJT for creativity, and design a corresponding DCE (Anderson, Thier, & Pitts, forthcoming). We found near-identical patterns of strengths, limitations, and trade-offs. Though we must note that the similarities across two metacognitive domains does not provide sufficient evidence for generalizability of our framework, it seems to be useful for self-report measures, which often leap into the assessment market because they are efficient to build, administer, scale, and validate. The challenge remains to compare between and within measurement approaches.

Because SJTs require respondents to weigh possible choices and use their best judgments, often in socially sensitive scenarios, item writing, especially for constructed-response items, requires high

degrees of sensitivity to attenuate the likelihood of self-presentation bias. In particular, SJTs provide a strong addition to a multi-method approach, testing a metacognitive skill in highly specified contexts (Soland et al., 2013). Concerns over self-presentation bias live with DCE construction and administration, as well. But careful construction, testing, and iterative revising has proven successful at reducing biases (Asplund, Lopez, Hidges, & Harter, 2009; Roberts et al., 2015). Our conceptual DCE provides a global citizenship example for K-12 use, but we do not represent the first attempt to use DCE, or a modified forced-choice format, to estimate skills or dispositions. The Clifton StrengthFinder® has been used with adults (Asplund et al., 2009). In K-12, Dweck's Implicit Personality Theory scale, has been adapted multiple times to create items that measure learners' motivational orientations (Beckmann, Beckmann, & Elliott, 2009; Ziegler & Stoeger, 2010). Given some promising evidence of reliability and validity, StrengthFinder® provides another framework worth considering when developing metacognitive assessments for K-12 education.

In addition to offering less-biased estimates, DCEs could also be constructed to measure environmental and instructional factors that relate significantly to metacognitive skills. As a result, practitioners and researchers could investigate an individual's perceived development, as well as the contextual factors that either augment or diminish this self-perception. Ultimately, DCEs can represent real-life scenarios more authentically. As a result, SJTs and DCEs might prove to be more sensitive to changes in students' perceptions, behaviors, and attitudes than self-reports, however, we did not explicitly include sensitivity to growth as a dimension of our evaluative framework despite its close relation to our included dimensions.

Finally, Soland et al. (2013) suggested examining the utility and feasibility of metacognitive skill measures in terms of their abilities to (a) provide actionable information to students and educators; (b) engage students in authentic contexts; and (c) encourage effective teaching and learning. Though the current study focused on aspects of measurement, we recognize the overriding importance to prioritize effective teaching and learning. To gain traction in the age of accountability that pervades many

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

developed nations' educational systems, appropriate and effective measures of metacognitive skills needs a reinvigorated effort.

Author Name: Michael Thier
Contact Email: mthier@uoregon.edu

References

Ahmed, W., van der Werf, G., Minnaert, A., & Kuyper, H. (2010) Students' daily emotions in the classroom: Intra-individual variability and appraisal correlates. *British Journal of Educational Psychology, 80,* 583-597.

Aronson, J., & Dee, T. (2012). Stereotype threat in the real world. *Stereotype threat: Theory, processes, and application*, 264-279.

Asplund, J., Lopez, S., Hodges, T., & Harter, J. (2009). *The Clifton StrengthsFinder 2.0 Technical report: Development and validation.* Gallup Consulting.

Aubusson, P., Burke, P., Schuck, S., Kearney, M., & Frischknecht, B. (2014). Teachers choosing rich tasks: The moderating impact of technology on student learning, enjoyment, and preparation. *Educational Researcher, 43,* 219-229.

Beckmann, N., Beckmann, J.F., & Elliott, J.G. (2009). Self-confidence and performance goal orientation interactively predict performance in a reasoning test with accuracy feedback. *Learning and Individual Difference, 19*(2), 277-282.

Brustein, W. (2009). It takes an entire institution: A blueprint for the global university. *The handbook of practice and research in study abroad: Higher education and the quest for global citizenship*, 249-265.

Conley, D. T. (2013). Rethinking the notion of "noncognitive." *Education Week, 32*(18), 20-l2.

Conley, D. T. (2015). A new era for educational assessment. *Education Policy Analysis Archives*, *23*, 8. http://dx.doi.org/10.14507/epaa.v23.1983

Conley, D. T., Beach, P., Thier, M., Lench, S. C., & Chadwick, K. L. (2014, June). *Measures for a college and career indicator: Innovative measures*. Eugene, OR: Educational Policy Improvement Center.

Cunningham, C. E., Deal, K., Rimas, H., Buchanan, D. H., Gold, M., Sdao-Jarvie, K., & Boyle, M. (2008) Modeling the information preferences of parents of children with mental health problems: A discrete choice conjoint experiment. *Journal of Abnormal Child Psychology, 36*, 1123-1138.

Cunningham, C.E., Vaillancourt, T., Rimas, H., Deal, K., Cunningham, L., Short, K., & Chen, Y. (2009) Modeling the bullying prevention program preferences of educators: A discrete choice conjoint experiment. *Journal of Abnormal Child Psychology*, *37*, 929-943.

Deardorff, D. K. (2014, May 15). Some thoughts on assessing intercultural competence.[Weblog post]. Accessed at https://illinois.edu/blog/view/915/113048 on Aug. 15, 2015

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087.

Duckworth, A. & Yaeger, D. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), *237-251.*

Duncan, A. (2013, November). *Building a Stronger Pipeline of Globally-Competent Citizens*. Speech presented at the international education week "Mapping the Nation": Making the case for global competency" Launch Event, Washington, DC.

Dweck, C. (2010). Even geniuses work hard. *Educational Leadership, 68*(1), 16–20.

Eidoo, S., Ingram, L. A., MacDonald, A., Nabavi, M., Pashby, K., & Stille, S. (2012). Through the kaleidoscope: Intersections between theoretical perspectives and classroom implications in critical global citizenship education. *Canadian Journal of Education*, *34*(4), 59-85.

Evans-Brown, S. (2015, March 5). Four N.H. school districts allowed to cut back on statewide standardized tests. *New Hampshire Public Radio.* Accessed at http://nhpr.org/post/four-nh-school-districts-allowed-cut-back-statewide-standardized-tests on Aug. 15, 2015

Fantini, A. E. (2009). Assessing intercultural competence. D. Deardorff (Ed.). *The SAGE handbook of intercultural competence*, 456-476.

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. Chicago: University of

Chicago Consortium on Chicago School Research.
Hansen, T. K. (2013). The Danish Simulator-Exploring the Cost-Cutting Potential of Computer Games in Language Learning. In proceedings: *ICT for language learning* (p. 271).
Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics, 54*(1), 3-56.
Hett, J. E. (1993). *The development of an instrument to measure global-mindedness* (Doctoral dissertation, University of San Diego). Retrieved from http://intl-jsi.sagepub.com/content/15/5/445.refs.
Howard, K. E., & Anderson, K. A. (2010) Stereotype threat in middle school: The effects of prior performance on expectancy and test performance. *Middle Grades Research Journal, 5*(3), 119-137.
Hunter, W. D. (2004). *Knowledge, skills, attitudes, and experiences necessary to become globally competent* (Doctoral dissertation, Lehigh University). Retrieved from http://www.globalcompetence.org/research/WDH-dissertation-2004.pdf.
Huws, N., Reddy, P. A., & Talcott, J. B. (2009). The effects of faking on non-cognitive predictors of academic performance in university students. *Learning and Individual Differences, 19*, 476-480.
Johnson, W. L., Friedland, L., Schrider, P., Valente, A., & Sheridan, S. (2011). *The Virtual Cultural Awareness Trainer (VCAT): Joint Knowledge Online's (JKO's) solution to the individual operational culture and language training gap*. In Proceedings of ITEC 2011. London: Clarion Events.
Johnson, W. L., & Zaker, S. B. (2012). *The power of social simulation for Chinese language teaching*. Proceedings of TCLT7. Honolulu, 2012.
Johnson, W.L. & Wu, S.M. (2008). *Assessing aptitude for learning with a serious game for foreign language and culture*. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), International Conference on Intelligent Tutoring Systems (p. 520-529). Berlin: Springer-Verlag.
Kennelly, B., Flannery, D., Considine, J., Doherty, E. & Hynes, S. (2014) Modelling the preferences of students for alternative assignment designs using discrete choice experiment methodology. *Practical Assessment, Research & Evaluation, 19(16).*
Killick, D. (2012). Seeing ourselves-in-the-world: Developing global citizenship through international mobility and campus community. *Journal of Studies in International Education, 16,* 372-389. DOI: 1028315311431893.
Kopcha, T., & Sullivan, H. (2007) Self-presentation bias in surveys of teachers' educational technology practices. *Education Tech Research Development, 55,* 627-646.
Lagattuta, K. H., Sayfan, L., & Bamford, C. (2012) Do you know how I feel? Parents underestimate worry and overestimate optimism compared to child self-report. *Journal of Experimental Child Psychology, 113,* 211-232.
Lai, E. R. (2011). *Collaboration: A literature review*. Pearson. Access on http:// ed.pearsonassessments.com/hai/images/tmrs/Collaboration-Review.pdf Aug. 15, 2015
Landine, J., & Stewart, J. (1998). Relationship between metacognition, motivation, locus of control, self-efficacy, and academic achievement. *Canadian Journal of Counselling, 32,* 200-212.
Lawthong, N. (2003). A development of the global-mindedness scale in Thai socio-cultural context. *Journal of Institutional Research South East Asia, 1*(2), 57-70.
Lievens, F., & Sackett, P. R. (2012) The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97,* 460-468.
Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-28.
Lombardi, A., Seburn, M., & Conley, D. (2011). Development and initial validation of a measure of academic behaviors associated with college and career readiness. *Journal of Career Assessment, 19,* 375-391.

Louviere, J. J., Islam, T., Wasi, N., Street, D., & Burgess, L. (2008) Designing discrete choice experiments: Do optimal designs come at a price? *Journal of Consumer Research, Inc. 35,* 360-375.

Merriam, S. B. (1998). *Qualitative Research and Case Study Applications in Education.* Revised and expanded from *"Case study research in education."* San Francisco: Jossey-Bass.

Miller, A. (2012) Investigating social desirability bias in student self report surveys. *Educational Research Quarterly, 36*(1).

Molina, S., & Lattimer, H. (2013). Defining global education. *Policy Futures in Education, 11*, 414-422.

Morais, D. B., & Ogden, A. C. (2011). Initial development and validation of the global citizenship scale. *Journal of Studies in International Education, 15*, 445-466.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st Century.* Committee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino and M.L. Hilton, Editors. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110 § 115, 1425 Stat. (2002)

Osterlind, S. J. (2009). Theory, principles, and applications of mental appraisal. Allyn & Bacon/Pearson.

Parkinson, A. (2009). The rationale for developing global competence. *Online Journal for Global Engineering Education, 4*(2), 1-15.

Partnership for 21st Century Skills (n.d.). List of exemplar schools. Accessed at http://www.p21.org/exemplar-program-case-studies/list-of-exemplar-schools) on Aug. 15, 2015.

Pecheone, R., Kahl, S., Hamma, J., & Jaquith, A. (2010). *Through a looking glass: Lessons learned and future directions for performance assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Reimers, F. (2010). Educating for global competency. In J. E. Cohen & M. B. Malin (Eds.). International perspective on the goals of universal basic and secondary education (pp. 183-2020). New York: Routledge.

Roberts, R., Martin, J., & Olaru, G. (2015). A Rosetta Stone for noncognitive skills: Understanding, assessing, and enhancing noncognitive skills in primary and secondary education. New York: Asia Society & Professional Examination Service.

Rothstein, R. (2004). Accountability for noncognitive skills: Society values traits not covered on academic tests, so why aren't they measured in school? *School Administrator, 61*(11), 29-33.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92-96.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Singh, M., & Qi, J. (2013). 21st century international mindedness: An exploratory study of its conceptualization and assessment. University of Western Sydney.

Soland, J., Stecher, B. M., & Hamilton, L. S. (2013). Measuring 21st-century competencies: Guidance for educators. New York: Asia Society & RAND.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

Taylor, C.M., Vehorn, A., Noble, H., Weitlauf, A.S., & Warren, Z.E. (2014) Brief report: Can metrics of reporting bias enhance early autism screening measures? *Journal of Autism Development Disorders, 44*, 2375-2380.

Thier, M. (2015, November). Left behind: School poverty, remoteness, and opportunity to learn global competence. Paper to be presented at the annual conference of the Australian Association of Research for Education.

Thier, M. (in press). Globally speaking: Cultural intelligences and cross-cultural competencies. In Y. Zhao (Ed.), *Counting what counts: Reframing educational evaluation*. Bloomington, IN: Solution Tree.

Thorndike, R. M., & Thorndike—Christ, T.M. (2010). *Measurement and evaluation in education and psychology (8th ed.).* Boston, MA: Allyn.

Vockell, E. (2009). Metacognitive skills. *Educational psychology: A practical approach :(Online Ed.)* Retrieved from *http://educationcalumetpurdue.edu/vockell/edPsybook/Edpsy7/edpsy7_meta.htm*.

Vrugt, A., & Oort, F. J. (2008). Metacognition, achievement goals, study strategies and academic achievement: pathways to achievement. *Metacognition and Learning, 3*(2), 123-146.

Waschbusch, D.A., Cunningham, C.E., Pelham, Jr., W.E., Rimas, H. L., Greiner, A.R., Gnagy, E.M., Waxmonsky, J., Fabiano, G.A., Robb, J.A., Burrows-MacLean, L., & Scime, M. (2011) A discrete choice conjoint experiment to evaluate parent preferences for treatment of young, medication naive children with ADHD. *Journal of Clinical Child & Adolescent Psychology, 40*, 546-561.

Waters, E., & Sroufe, L. A. (1983). Social competence as a developmental construct. *Developmental Review, 3*(1), 79-97.

Weijters, B., Geuens, M., & Schillewaert, N. (2010) The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*(2), 105-121.

Willard, J. (2009). Global competency. Language Corps. Accessed at http://www.nafsa.org/_/file/_/global_competency_2.pdf on Aug. 15, 2015.

Zhao, Y. (2010). Preparing globally competent teachers: A new imperative for teacher education. *Journal of Teacher Education, 61*, 422-431.

Zhao, Y. (2012). *World Class Learners*. Thousand Oaks, CA: Corwin.

Zhao, Y. (in press). *Counting what counts: Reframing educational evaluation*. Bloomington, IN: Solution Tree.

Ziegler, A. & Stoeger, H. (2010). Research on a modified framework of implicit personality theories. *Learning and Individual Differences, 20,* 318-326.