Learning and Individual Differences xxx (2016) xxx-xxx

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com

# Interpersonal and intrapersonal skill assessment alternatives: Self-reports, situational judgment tests, and discrete choice experiments☆

Ross Anderson *, Michael Thier, Christine Pitts

*Educational Policy Improvement Center, 1700 Millrace, Eugene, OR 97403, USA*

ARTICLE INFO

*Article history:*
Received 26 September 2015
Received in revised form 17 September 2016
Accepted 23 October 2016
Available online xxx

*Keywords:*
Creative thinking
Discrete-choice experiments
Global citizenship
Self-report biases
Situational judgment tests

ABSTRACT

Responding to a groundswell of researcher and practitioner interest in developing students' interpersonal and intrapersonal skills, we evaluated three measurement approaches for creativity and global citizenship. We designed a 10-criteria evaluative framework from seminal and cutting-edge research to compare extant self-reports and situational judgment tests (SJTs) from each construct and to design two discrete choice experiments (DCEs). Our evaluation detailed opportunities, challenges, and tradeoffs presented by each approach's design considerations, possibilities for bias, and validity-related issues. We found that researchers rely heavily upon self-report instruments to measure constructs, such as creative thinking and global citizenship. We found evidence that the self-report instruments evaluated were susceptible to some biases more than others. We found that SJTs and DCEs may mitigate some concerns of bias and validity present in self-report when measuring interpersonal and intrapersonal skills. We make recommendations for future development of these formats.



## 1. Introduction

Responding to the Every Student Succeeds Act (ESSA), states' new systems of accountability will include "not less than one indicator of school quality or student success" (Every Student Succeeds Act, 2015, p. 35). To operationalize school quality or student success, ESSA offers vague examples of complex constructs such as student and/or educator engagement, school climate and safety, and postsecondary readiness. Importantly, ESSA allows states to select "any other indicator the State chooses that meets the requirements of this clause" (ESSA, 2015, p. 35). Such flexibility invites states' unique interpretations of the many educational opportunities that may support quality schools and successful students.

ESSA tacitly encourages states to focus on the wide array of interpersonal and intrapersonal skills that decades of research indicate as essential for student success in college, career, citizenship, and building a fulfilling life (National Research Council, 2012). Interpersonal and intrapersonal skills (a) span academic disciplines; (b) may be more transferrable and applicable for 21st-century students than the highly esteemed cognitive skills gained through instruction in mathematics and reading (National Research Council, 2012); (c) are at least equal to cognitive skills in their ability to predict postsecondary success (Conley & Darling-Hammond, 2013); (d) are more malleable than cognitive skills (Heckman, 2000); and (e) predict long-term academic and economic outcomes (Soland, Stecher, & Hamilton, 2013). Yet, measures of such skills are neither fit to inform classroom-level decisions nor do they serve accountability purposes (Duckworth & Yeager, 2015). The current study investigates promising approaches that need greater attention to improve research, support schools, and shift accountability priorities.

### 1.1. K–12 priorities

Both research findings and employers' calls to foster interpersonal and intrapersonal skills have leveraged some school systems and states to prioritize these skills through statutory requirements and instructional agendas. For instance, Maine graduates in the Class of 2019 will need to demonstrate proficiency in five *Guiding Principles* that capture a broad set of skills and dispositions important to college, career, and citizenship readiness. Maine expects its alumni to be integrative and informed thinkers, self-directed and lifelong learners, clear and effective communicators, responsible and involved citizens, and creative and practical problem solvers (Fukuda, Anderson, & Lench, 2015; Maine Department of Education, 2015).

Despite policy changes, the infancy of literature on cultivating interpersonal and intrapersonal skills across K-12 leaves opportunities for systemic innovations. Many interpersonal-intrapersonal domains overlap with conceptualizations that other states, districts, and organizations have created through industry and higher education partnerships. Still, in light of recent emphasis, few schools or districts measure interpersonal and intrapersonal skill development explicitly or prioritize their developmental trajectories (Conley, 2015; Farrington et al., 2012; Rothstein, 2004).

? Ross Anderson is Senior Lead Researcher and Michael Thier and Christine Pitts are Research and Policy Fellows at the Educational Policy Improvement Center in Eugene, Oregon.

* Corresponding author.
Email address: ross_anderson@epiconline.org (R. Anderson)

http://dx.doi.org/10.1016/j.lindif.2016.10.017
1041-6080/

*1.1.1. Creative thinking and global citizenship*

For this study, we selected two interpersonal and intrapersonal skills that have been included in numerous frameworks, including Maine's *Guiding Principles*: creative thinking and global citizenship. Although those constructs reflect terminology from education research, conventions vary across disciplines. We chose creative thinking as 1-of-2 constructs of interest because educators, business leaders, and scholars consider it essential to entrepreneurship and quality of life (Csikszentmihalyi, 1996; Sternberg, 2006; Wagner, 2012), innovation and national progress (Florida, 2002; Zhao, 2012), and meaningful learning across disciplines (Anderson, 2015; Beghetto, 2016). Problematically, researchers have identified downward trends, both developmentally and over past decades, in K-12 students' development of creative thinking habits (Engel, 2009; Kim, 2011).

Equity concerns prompted our inclusion of global citizenship, an articulated goal of K-12 and higher education (Morais & Ogden, 2011; Zhao, 2010). Gaps in students' opportunities to learn and develop skills and dispositions related to global citizenship are associated with numerous factors, including socioeconomics (Bunnell, 2009), race/ethnicity (Perna et al., 2013), geography (Provasnik et al., 2007; Thier, 2016), or other factors that privilege some students over others (Killick, 2011; Reimers, 2009). Importantly, demands from rapid globalization establish an impetus to provide educational experiences that enhance students' global citizenship development in order to prepare for economic and social futures that will require it (Duncan, 2013; Molina & Lattimer, 2013). Table 1 details our operationalizations of creative thinking and global citizenship alongside conceptualizations from Maine's *Guiding Principles*.

*1.1.2.*

School and district reluctance to commit to interpersonal and intrapersonal skill measurement is somewhat justifiable. The diffuse and overlapping definitions of interpersonal and intrapersonal skills—evident in Table 1—challenge psychometricians' abilities to design reliable measures (Soland et al., 2013). As the field awaits appropriate measures, practitioners and policymakers should take thoughtful steps to understand the nature of skill development and associated measurement challenges before entrenching these skills in the complexities of K-12 accountability (Duckworth & Yeager, 2015). Interpersonal and intrapersonal skill measurement to date has exposed concerns of measurement bias that are endemic to most self-reports (Roberts, Martin, & Olaru, 2015), the most common instrument type in social science research (Weijters, Geuens, & Schillewaert, 2010). Duckworth and Yeager (2015) explained that respondents follow a five-stage cognitive sequence when completing self-reports: (a) understanding the item, (b) recalling relevant information, (c) integrating one's existing context, (d) translating judgment into a response option, and (e) editing the response if necessary. Despite the prevalent use of self-reports to collect data in the social sciences (Weijters et al., 2010), various reliability/validity threats occur at each stage (Duckworth & Yeager, 2015).

For example, language and comprehension problems, difficulty integrating past and present behaviors, misinterpretation of student behaviors or meanings of scales, and other response biases can occur when administering self-reports, even when developers attempt to account for them (Duckworth & Yeager, 2015). In some cases, typical self-report ratings can trigger egocentric, faking, or self-presentation biases. Responses, therefore, may artificially exaggerate results, yielding inaccurate evaluation (Huws, Reddy, & Talcott, 2009; Kopcha & Sullivan, 2007; Lagattuta, Sayfan, & Bamford, 2012). Associating stakes with respondents' scores on measures of skills or dispositions may increase potential for biases (Duckworth & Yeager, 2015). Such potential supports the rationale for privileging measurement of, therefore esteeming, cognitive domains such as literacy and numeracy (Conley & Darling-Hammond, 2013). Consequently, stakeholders have missed opportunities to identify achievement gaps that could be functions of student effort and/or dispositions rather than sheer aptitude (Pecheone, Kahl, Hamma, & Jaquith, 2010). Measuring interpersonal and intrapersonal capabilities may be a particularly powerful support to practitioners' pedagogy to account for students' assets and that respond to their needs.

Although challenges of rigor and quality abound, interpersonal and intrapersonal skills do not suffer from a lack of measures. Soland et al. (2013) notes that their exclusion from the testing industry's robust focus on cognitive skills denied opportunities to incorporate the interdependence of different types of skills into assessment design. Enduring from the 1950s, standardized cognitive tests have become a cultural norm in many high-income countries, yielding the prevailing sentiment that interpersonal and intrapersonal skill assessments are not rigorous (Conley, 2015). To confront this pervasive belief, the current study compared three measurement approaches: self-reports, situational-judgment tests, and discrete-choice experiments. We reviewed technical issues and illustrated strengths, limitations, and tradeoffs of the alternatives to self-reports. Though self-reports build on a century of use across academic disciplines, the newer approaches have emerged recently in education as opportunities to enliven research, practice, and policy.

**Table 1**
Definitions of interpersonal and intrapersonal skills included in this study.

| Skill | Definition | Maine guiding principle (Fukuda et al., 2015) |
|---|---|---|
| Creative thinking | 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) | *Creative and Practical Problem Solver*: Exploring and formulating; cultivating and selecting ideas; taking risks and tolerating ambiguity; and validating with others and reflecting on learning |
| Global citizenship | Awareness of global issues, particularly those related to social justice; knowledgeable about global interdependence and cultural processes within many nations; able to take multiple perspectives and communicate across languages and cultures; and values diversity of people and beliefs, plus feels empathy toward/respect for others and a responsibility to act (Thier, Thomas, Tanaka, & Minami, 2016) | *Responsible and Involved Citizen*: Affinity, belonging, and ownership in a community; engaging with and valuing diverse perspectives; participating in civil discourse and collaborative decision making; and understanding and acting on issues of local, regional, and global significance. |

*Note*. Thier et al. (2016) offer the most recent global citizenship definition amid a field where definitions remain contested. Türken and Rudmin (2013) characterize concepts such as globalism, cosmopolitanism, multiculturalism, internationalism, transnationalism, worldism, worldmindedness, and glocalization as constructs with overlapping meanings. Singh and Qi (2013) add global citizenship to such lists of conceptually overlapping constructs, such as common humanity, cultural intelligence, global competence, global mindedness, intercultural understanding, international mindedness, multiliteracies, omniculturalism, and peace and development.

*1.2. Definitions*

In the subsequent section, we describe the framework we developed to evaluate extant *self-reports* and *situational-judgment tests* (SJT). Using our framework, we evaluated four measures: a self-report and a SJT each to measure respondents' creative thinking and

global citizenship, respectively. Our framework also guided development of *discrete-choice experiments* (DCE) for each construct of interest. Typical examples of self-reports include questionnaires, in which respondents—independent of a researcher—read and respond to a statement by selecting a response option. Constructed with hypothetical scenarios, SJTs can measure procedural knowledge or skills within specific domains, replacing formal field observations. SJTs can measure decision-making, problem-solving, interpersonal, and organizational skills (Lievens & Sackett, 2012). SJTs may provide an optimal approach to assessing situational skills by creating relevant scenarios for learning contexts. SJTs can use multiple choice, constructed responses, rankings, or Likert-style ratings as response options and can incorporate various multimedia formats (Roberts et al., 2015).

From simple paired comparisons to challenging choices with multiple dimensions and attributes, DCEs can feature varying levels of complexity and cognitive loads (Aubusson, Burke, Schuck, Kearney, & Frischknecht, 2014; Roberts et al., 2015). Presented with hypothetical choices that vary on descriptive dimensions within a choice set, respondents state preferences via ranking or selection. Researchers model respondents' preferences statistically based on these decisions to estimate the values (i.e., weights) of attributes that contribute to their choices (Kennelly, Flannery, Considine, Doherty, & Hynes, 2014). In its simplest form, DCEs consist of two related choices of behavior, attribute, attitude, or disposition that represent contrasting levels of development. Use of these *forced-choice* measures with K–12 students has been limited with some application for social/emotional skills, attitudes, and behaviors (e.g., Nett, Goetz, & Daniels, 2010; Lau & Roeser, 2008; Beckmann, Beckmann, & Elliott, 2009). Forced-choice formats may reduce measures' susceptibility to faking and social desirability, potentially providing more valid representations of performance or development (Cao, 2016; Drasgow et al., 2012; Jackson, Wroblewski, & Ashton, 2000).

## 2. Method

To establish the literature pool for this review, we searched ERIC for peer-reviewed articles from 2006 to 2015 with "bias" and the following keywords/variants: self-report, situational-judgment test, and discrete-choice experiment. We found 206 articles: 95 for self-report, 81 for SJT, and 30 for DCE. We refined our pool to only articles that had titles, keywords, or descriptors with measurement, validity, and/or reliability, yielding 15 studies. We added one current study that illustrated a DCE in this paper's educational scope (Aubusson et al., 2014). We synthesized the 16 qualifying studies to develop a framework for evaluating strengths, limitations, and tradeoffs of the three measurement approaches. Our review informed an evaluative framework with three dimensions: (a) measurement design, (b) possibilities for bias, and (c) validity. Using this framework, we evaluated creative thinking and global citizenship measures based on their potential use with U.S. high school students.

### *2.1. Framework for evaluating measurement approaches*

As we show in Fig. 1, the innermost circle of our evaluative framework features *measurement design* because it relates most directly to the primary artifact of evidence: the measurement tool. The second dimension, *possibilities for response bias*, encompasses considerations beyond the tool itself, such as respondents' interactions with the tool. As the final dimension, *validity* relates to the many possible interpretations, both intended and unintended, that might arise from administering the tool.

#### *2.1.1. Measurement design*

This dimension focuses on decisions that test creators made, reported, or neglected to make-report during measure development. We examined *preparatory qualitative measures*, *measures of internal consistency*, and *efficiency tradeoffs*, all of which point toward model specification, item construction, reliability, and design decisions.

#### *2.1.2. Preparatory qualitative measures*

Informing design through qualitative measures ensures broader alignment between respondent populations and test items, which may increase the validity of findings. Test designers' a priori assumptions might limit the factors they include in measures (Cunningham et al., 2009). Qualitative data collection can guide construction, wording, or item order (Cunningham et al., 2009; Kennelly et al., 2014).

#### *2.1.3. Measures of internal consistency*

Test designers traditionally rely on internal consistency measures, such as item correlations with analyses such as Cronbach's alpha, split-half, and Kuder-Richardson (Osterlind, 2009). In other cases, researchers use survey methodology (questionnaires and interviews), choice experiments, and focus groups to decrease rates of false positives (Taylor, Vehorn, Noble, Weitlauf, & Warren, 2014). Taylor et al. (2014) used internal metrics of response characteristics to analyze possible biases in instances where multiple measures are infeasible. In the case of DCEs, Kennelly et al. (2014) recommended using status quo options for respondents, so they are not forced into making uninformed choices when encountering unfamiliar items.

#### *2.1.4. Efficiency tradeoffs*

If employing measures that require respondents to choose, compare, or judge between choices, test designers should negotiate tradeoffs between statistical efficiency and cognitive load. SJTs measure novel tasks that are demanding if they represent new experiences for respondents. A *forced-choice* approach changes a DCE method by limiting choice sets' attribute amounts; however, DCEs require multiple choice sets to achieve adequate reliability (Roberts et al., 2015). As a tradeoff, tasks require more cognitive demand when complexity increases in numbers of either attributes or choice sets. Moreover, DCEs might rely on computer software to control for the number of attributes or choice levels (Kennelly et al., 2014). In such cases, Louviere, Islam, Wasi, Street, and Burgess (2008) caution researchers to consider possible drawbacks: respondents often choose the least cognitively demanding answer resulting in response variability (i.e. error variance) possibly biasing estimates.

#### *2.1.5. Potential response bias*

The response-bias dimension focuses on factors that can influence item responses, including: *self-presentation, egocentric, stereotype threat,* and *response style.*

#### *2.1.6. Self-presentation*

Self-presentation bias presents itself when items produce "greater-than-actual" tendencies due to respondents' interpretations of socially desirable responses (Kopcha & Sullivan, 2007, p. 14), resulting in *faking* (Huws et al., 2009; Roberts et al., 2015). Self-reports are both commonly used and particularly susceptible to self-presentation (Duckworth & Yeager, 2015). Self-presentation (i.e., social desirability) bias is typically measured using socially sensitive subject matter only. Yet, Miller (2012) recommends that researchers explore self-presentation bias for all topics.

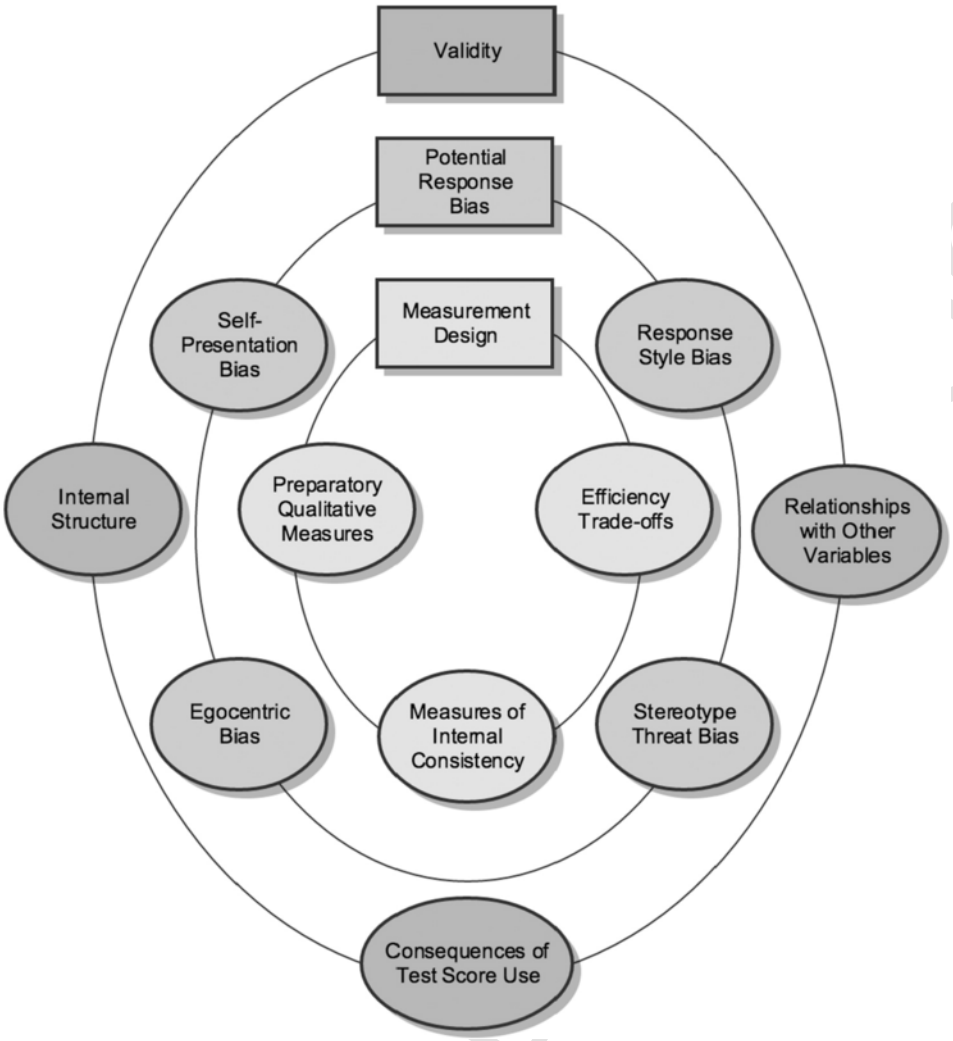


**Fig. 1.** This tripartite evaluative framework for measurement approaches includes 10 criteria to consider in terms of design, potential biases, and validity considerations.

*2.1.7. Egocentric bias*

In measures where respondents are likely to use their personal views to determine response choices, personal context should be considered as a possible source of bias for individual or aggregated responses (Cunningham et al., 2008; Lagattuta et al., 2012). In one case, Lagattuta and colleagues found a convergence of child- and parent-reported emotions. As parents' ratings of their own worries or optimism increased, *their* ratings of *their child's* worries or optimisms also increased. In DCEs, choice options can illustrate a small portion of the real-world decisions that respondents make (Cunningham et al., 2008). Therefore, respondents' varying backgrounds contextualize responses and might create bias that researchers must consider when modeling results (Waschbusch et al., 2011). Ultimately, refined sampling procedures and preparatory qualitative inquiry might reduce susceptibility to egocentric bias.

*2.1.8. Stereotype threat*

Stereotype threat, the risk of confirming a negative stereotype about one's group (Steele & Aronson, 1995), can bias findings if respondents' appraisals associate with their emotions that connect to stereotypes. Judgments about respondents' performances on new tasks are more susceptible to stereotype threat when compared to those situated within their natural setting (Howard & Anderson, 2010). Ahmed, Van der Werf, Minnaert, and Kuyper (2010) examined emotions and appraisals of a math lesson, finding students to experience emotions and associate them directly with their self-reported competence of the lesson and their perceptions of the lesson's value. However, using SJTs, Howard and Anderson (2010) simultaneously stigmatized students' race-ethnicity and examined student performance on novel tasks. Respondents reported negative effects on their expected performances; no effect on actual performances was detected. Therefore, stereotype threat might affect measured self-perception of future performance, such as responses on pre-employment surveys (Lievens & Patterson, 2011).

*2.1.9. Response-style bias*

Response-style bias is measured with a systematic assessment of inconsistencies within responses over multiple items in a single measure (Weijters et al., 2010). Inconsistencies in responses are interpreted as unsystematic; most researchers regard them as random error. However, response-style biases occur when measures disproportionately and systematically produce positive (e.g., acquiescence-response style) and/or extreme responses. Such patterns are not attributed to random error. Instead, Weijters et al. (2010) recommend cap-

turing variance via reverse-coded items or homogenous and heterogeneous content to diagnose and correct systematic response-style bias.

### 2.2. Validity

The validity dimension focuses on the degree to which evidence, theory, and test score interpretations align with the proposed use of the test score (Joint Committee on Standards for Educational and Psychological Testing, 2014)[1]. We examined evidence of *internal structures*, *consequences of test score use,* and *relations to other variables*.

#### 2.2.1. Internal structure

The Joint Committee on Standards for Educational and Psychological Testing (JCSEPT, 2014) expect researchers to analyze the degree to which tests' items and their components relate to the construct(s) upon which they base test scores. In addition, researchers should consider that evidence of tests' internal structures depends upon various factors, including research designs and sampling strategies, when analyzing data from measures. In some cases, sampling specificity limits the evidence gathered to support the internal structure of the measure (Waschbusch et al., 2011). In addition, the measurement approaches we consider in the current study—self-reports, SJTs, and DCEs—all measure single time points of respondents' perspectives and behaviors, which cannot capture the evolving complexities of attitudes, characteristics, or contexts (Cunningham et al., 2009).

#### 2.2.2. Consequences of test score use

Validating test scores also involves gathering evidence to assess whether proposed interpretations of scores align with the scores' intended uses (JCSEPT, 2014). For DCEs, poor measurement design sets the stage for negative consequences of varying test score interpretations (Cunningham et al., 2009; Kennelly et al., 2014). Certain context-bound aspects of assessment, such as cost or value, result in varying inferences across respondents, troubling the inferential power of test scores (Kennelly et al., 2014). Overall, researchers must consider how test scores, (a) are interpreted based on test designers' intended uses, (b)are used for claims beyond the scope of the intended uses, and (c) might result in unintended consequences for subgroups of respondents (JCSEPT, 2014).

#### 2.2.3. Relations to other variables

Common intended interpretations for test scores include the analysis of how constructs relate to other variables. Therefore, when measures forecast performance in other criteria, evidence supports the relations of test scores to external variables (JCSEPT, 2014). Traditionally, self-reports and SJTs have aided decisions regarding job or school placements. SJTs predicted job and internship performance better than self-reports and other cognitive factors, supporting their interpretative potential to provide information about relations to external variables (Lievens & Sackett, 2012). An inherent threat to gathering predictive evidence resides in measuring novel tasks with SJTs, because they are susceptible to egocentric and stereotype biases (Howard & Anderson, 2010). Although Lievens and Sackett (2012) recommend using SJTs to measure skills related to students' internal beliefs, SJTs are susceptible to measurement design biases that could be exacerbated when analyses weight scenarios and options differentially (Howard & Anderson, 2010).

[1] The Joint Committee on Standards for Educational and Psychological Testing (JCSEPT) reflects the work of the *American Educational Research Association*, *American Psychological Association*, and the *National Council on Measurement in Education*.

## 3. Results

In this section, we present results using our evaluative framework to analyze the reliability, validation, and descriptive studies of extant self-report scales and SJTs for creative thinking and global citizenship. We also demonstrate conceptual designs of DCEs for each construct of interest. Runco, Plucker, and Lim (2001) developed our first self-report, the Runco Ideational Behavior Scale (RIBS), which includes 5-point, Likert-type items ($n = 23$) to measure the openness, fluency, and originality of individuals' creative thinking processes by targeting ideational behaviors, attitudes, and self-belief. Our second self-report, the Global Identity Scale (GIS; Türken & Rudmin, 2013), includes 6-point, Likert-type items ($n = 10$) designed to measure global identity and have cross-cultural applicability. Türken and Rudmin purport that the GIS taps into cultural openness and non-nationalism.

SJTs have gained more attention recently as an approach to measurement. Therefore, we employed our framework to organize any available published findings on SJTs, but also apply the framework as an interpretative lens to discuss dimensions upon which literature remains sparse. Some of the most commonly used SJT-type assessments in creative thinking, such as the Many Uses Game or Many Instances Test, date back 50 years (Wallach & Kogan, 1965). We chose the Social Games assessment, part of the Runco Creativity Assessment Battery (rCAB; Runco, 2014), because the Many Instances Test lacks publically available information and the situations presented to a respondent are purposefully general and vague. Also, the rCAB Social Games combines two critical dimensions of creative thinking—originality and appropriateness. Respondents gauge appropriateness of their creative ideas as they navigate familiar, delicate social situations. Though Runco does not explicitly call Social Games an SJT, the measure fits the definition. Similarly, the computer-based simulation, Virtual Cultural Awareness Trainer (VCAT) requires that users demonstrate capacities in intercultural communication, problem solving, and cultural knowledge, among others (Johnson, Friedland, Schrider, Valente, & Sheridan, 2011). Likewise, VCAT's designers do not call their product a SJT of global citizenship, but their interactive role-playing scenarios require users decide how to respond "to different types of unfamiliar, desirable, or undesirable reactions from non-player characters" (p. 5–6). Both measures represent innovations from the typical multiple-choice SJT (Roberts et al., 2015).

Finally, as DCEs first entered educational research recently (see Aubusson et al., 2014), no validation studies to date have interrogated DCEs for the constructs of interest. Rather, we constructed DCE exemplars to illustrate the potential benefits, challenges, and tradeoffs, as well as the design decisions required to operationalize and measure interpersonal and intrapersonal skills. Various sectors have used DCEs to learn about the predictive power and nuances of consumer preferences and to understand dispositional differences of current and future employees (Cao, 2016). Considering students are the most important consumers of educational policies and practices, DCEs could assess educational opportunities in many ways. Part of this study includes the examination of DCEs' potential to approximate individuals' interpersonal and intrapersonal skills as consumers of learning.

### 3.1. Self-report: Creative thinking

In Table 2, we report strengths, limitations, and tradeoffs for the RIBS creative thinking measure and the GIS, a proxy for global citizenship.

#### 3.1.1. Measurement design

Runco et al. (2014) do not report using *preparatory qualitative measures* to inform their design of the RIBS. Rather, they refer to creativity theory and the body of extant measures that use different approaches to target creative potential. Authors cite literature that reflects poor evidence supporting relations between test score use and external variables for most prior tests of creative potential (Wallach & Kogan, 1965; Runco, 1986), suggesting that predictors might not be at issue, but rather inappropriate choices of criteria. As such, they position RIBS as an alternative criterion, one that captures the behaviors and attitudes toward ideation that approximate facets that divergent-thinking tests measure: originality, fluency, and flexibility (Runco et al., 2001). In one study, Plucker, Runco, and Lim (2006) found scores on a typical divergent-thinking test to predict RIBS scores. Using confirmatory factor analyses, a strategy to test an assessment's theoretical framework and how items contribute to the measured construct, the RIBS' initial development compared the cohesion of items for responses from two college student samples (Runco et al., 2001), allowing authors to narrow the item count from 93 to 23. In addition to finding high *internal consistency* ($\alpha = 0.92$), the theorized one-factor solution fit Sample 1 ($n = 97$); two-correlated factor solution fit Sample 2 ($n = 224$).

#### 3.1.2. Potential response bias

Self-presentation bias may be less problematic when items target attitudes rather than behaviors or practices (Kopcha & Sullivan, 2007). On the original RIBS (Runco et al., 2001), 12-of-23 items focus on beliefs or attitudes from one's past or present self. Another 11 items target retrospective or future-oriented reflection of respondent behavior. Depending on whether ideational behavior is a personally sensitive topic for a given respondent, bias toward *self-presentation* might threaten RIBS scores; however, it seems plausible that this bias would be greater for a checklist of one's creative accomplishments (or lack thereof).

In regards to *egocentric bias*, 5 items in the original RIBS explicitly ask respondents to compare one's self to others (e.g., "I think about ideas more often than most people"). Lagattuta et al. (2012) found that respondents might unwittingly generate a reference bias in their own perceptions when response choices require them to draw upon their own unique contexts. With the RIBS, items do not specify contexts; respondents may select contexts on their own. If a referenced context could be conceived as a linear relation, this issue might not be so problematic. For instance, a highly acclaimed musician responds that she "sometimes" combines ideas that others haven't, using her highly accomplished peers as a reference group for "others." A disengaged high school student might respond "often" comparing himself to his group of friends. The validity of comparisons between the two becomes tenuous.

As a standalone assessment, the RIBS does not present *stereotype threat*. However, authors intended the RIBS to serve within a battery of assessments that would depict creative potential (Runco et al., 2014). If a RIBS respondent were also expected to take a test for divergent thinking, an individual who self-rated low on the ideational behavior scale might enter the divergent-thinking test with lower self-expectations, feel less motivated, and be less likely to remain open to the creative task. The Expectancy-Value Model (Eccles & Wigfield, 2002) suggests that respondent internalizing of lower expectations will impact performance and, in this case, reinforce negative beliefs that some people aren't creative, a myth that may lower self-perceptions of creative potential.

Unless researchers use internal metrics to track response patterns and avoid response sets, *response-style bias* can arise. To boost reliability, Runco et al.'s (2014) incremental validity study added three new scales as internal metrics to the RIBS—a 12-item lie scale (items that everybody should respond to the same way), a 7-item contraindicative scale (items that represent the opposite or absence of the construct), and a 13-item distractor scale (items that are theoretically unrelated). Runco and colleagues found nonsignificant relations between the RIBS and both the distractor and contraindicative scales, but a large correlation with the lie scale and the RIBS. Though the new scales did not improve concurrent validity, the authors found potential utility for the new items on "a behavioral level, keeping respondents on their toes and mindful" (p. 196)—one explicit *efficiency trade-off*.

#### 3.1.3. Validity

If this measure is to be used for a high school population, future work must identify additional evidence in terms of both internal structure and predictive validity for future creative achievement. Runco et al. (2014) developed the RIBS to fill a creativity research gap, which lacked robust criteria to investigate temporal relations between tests of creative potential and actual development and fulfillment of that potential. Research documenting the degree to which RIBS test score interpretations are aligned with the intended use is limited; thus, *consequences of score interpretations* are left unknown. For instance, given the current body of research on the RIBS inclusion in a battery to screen for gifted and talented programs does not seem appropriate.

Though we did not find studies reporting on the evidence of relations between test scores and distal outcomes, studies of *relationships with other variables* found evidence of concurrent validity ($r = 0.47$, $p < 0.001$) with the measure of Creative Activity and Achievement Checklist (Runco et al., 2014) and one theoretically related scale and one theoretically divergent subscale of Basadur's behavior scale (Runco et al., 2001) and discriminant validity (no correlation detected) between RIBS scores and college grade-point averages (Runco et al., 2001). RIBS scores have been associated with more successful entrepreneurship (Ames & Runco, 2005), fluid intelligence, happiness, locus of control (Pannells & Claxton, 2008), and

**Table 2**
Strengths, limitations, & tradeoffs of self-reports for creative thinking and global citizenship.

| Measure | Strengths | Limitations | Tradeoffs |
|---|---|---|---|
| Runco Ideational Behavior Scale (2001) | Thorough measurement design; reliability; concurrent validity; balances behavioral and attitudinal items; shows potential use as a criterion | Reference bias; stereotype threat depending on administration; no evidence of predictive validity | Internal metrics vs. short form; No use of anchoring vignettes to measure reference bias; only college student (convenience) sample studied |
| Global Identity Scale (2013) | Exhaustive audit trail; attempts to mitigate cognitive load burden | Potential for Westernized egocentric bias and social desirability; Scale operationalization leads to response-style concerns | Expanding users to include high school students could compromise moderate-to-strong measures of internal consistency; Mixed evidence of concurrent validity |

higher levels of openness to experience and lower levels of conscientiousness (Batey, Chamorro-Premuzic, & Furnham, 2010). In total, these findings show promise for the RIBS' concurrent validity.

### 3.2. Self-report: Global citizenship

#### 3.2.1. Measurement design

Türken and Rudmin's thorough audit trail drew recognition for how they designed the GIS (Ozer & Schwartz, 2016). Even Sheehy-Skeffington's (2013) critique praised the GIS's "theoretical complexity and empirical rigor" (p. 90). Following DeVellis' (2003) scale-development process, Türken and Rudmin report (a) an exhaustive literature review; (b) open-ended questionnaire interviews of global identity definitions from 137 university students across 24 countries; (c) 392 items from 14 extant measures of related constructs; (d) item-performance ratings from 6 social scientists; (e) responses to 110 quantitative items from 1695 university students in Norway ($n = 684$), Turkey ($n = 605$), and the United States ($n = 406$); and (f) factor analyses of 8-, 10-, and 12-item versions of the GIS. Türken and Rudmin searched 15 databases with various keywords[2] to trace global identity conceptualizations from Ancient Greece, Rome, and Egypt through the European Enlightenment to modern depictions due to globalization. *Preparatory qualitative measures* from student interviews generated 21 new items, and authors removed 19 items from consideration because 5 or more respondents indicated their ambiguity, confusion, or poor expression.

GIS's *measures of internal consistency* meet or exceed generally recognized thresholds: Cronbach's alphas were reasonably high for scores from Norway ($\alpha = 0.79$), Turkey ($\alpha = 0.81$), and the United States ($\alpha = 0.85$). By contrast, the GIS produced mixed results in other studies.[3] Confirmatory factor analyses revealed two factors, which Türken and Rudmin dubbed cultural openness and non-nationalism. Both conform to their definition of global identity: "willingness to engage with the cultural *other* in a positive way" (p. 71, authors' emphasis). Türken and Rudmin did not report goodness-of-fit indices, but others' cross-cultural analyses confirmed the factor structure with Indian (Ozer & Schwartz, 2016) and Nigerian samples (Nwafor et al., 2016). Regarding *efficiency tradeoffs*, Türken and Rudmin chose the 10-item GIS because it performed equally well to the 12-item version. Both improved upon the 14 measures Türken and Rudmin's used to establish their item pool, where $M = 28.00$ items ($SD = 26.02$).

#### 3.2.2. Potential response bias

Despite in-depth reportage of design procedures, Türken and Rudmin offered less clarity about GIS's resistance to biases. For example, the authors (a) administered Strahan and Gerbasi's (1972) Social Desirability Scale (SDS) alongside the GIS and (b) changed early GIS items (e.g., "I consider myself a citizen of the world" to "I consider myself more as a citizen of the world than a citizen of some nation"), efforts to mitigate *self-presentation bias*. However, SDS scores produced marginally acceptable reliability ($\alpha = 0.70$) and correlated positively with Turkish ($r = 0.09$, $p < 0.05$) and U.S. respondents' GIS scores ($r = 0.15$, $p < 0.05$). Social desirability did not associate with Norwegian scores. Still, Türken and Rudmin recommend GIS users to employ covariates that control for social desirability. Sheehy-Skeffington suggests expanding the GIS to include less desirable global identity traits (e.g., confusion, rootlessness, underappreciation of local attributes) to curtail social desirability.

Sheehy-Skeffington noted another GIS limitation: convenience sampling led to its "narrow cross-section" of "mostly White, middle class college students … in the post industrial West" (p. 90). Therefore, respondents whose viewpoints do not align to Western norms might bias results. As such, GIS administration could yield *egocentric bias* and/or *stereotype threat*, particularly among K-12 students if they do not identify with cultural norms often codified in U.S. schools. Particularly one item leading respondents to evaluate their connection to "my own country" and another to consider "my own culture" might create cultural dissonance for students in marginalized groups. Furthermore, authors forward- and back-translated the GIS from English into Turkish, but they split the Turkish sample linguistically. Turkish students of English ($n = 144$) took the GIS in English; all other Turkish students ($n = 461$) took the Turkish version.

Even though authors randomized items to avoid order effects, *response-style bias* might present the GIS' most pressing problem. Five negative-keyed items form the non-nationalism subscale, which might simply tap nationalism or patriotism, not global identity's inverse. Türken and Rudmin suggest the positive/negative split of their scale reflects an approach-avoidance dynamic found frequently in psychological theories. Instead, Sheehy-Skeffington highlights the difficulty of knowing what the GIS measures: positive/negative perspectives on global identity or "a set of 'nice' versus 'nasty' sounding items" (p. 92).

#### 3.2.3. Validity

GIS' *internal structure* presents possibilities and challenges: Türken and Rudmin used 22 psychometric criteria during development[4] with both a "theory-driven, top-down approach and an empirically driven, bottom-up approach" (p. 84), but note their conflation of attitudes and identifications. Regarding *consequences of test score use*, the cross-cultural correlations of factor loadings ($r > 0.94$ for each national pairing) suggest suitability for international comparative research. Cautiously, though, Türken and Rudmin called for GIS administrators to measure covariates (e.g., age, sex, and education) because they expect cultural variation. Furthermore, Türken and Rudmin recognize gaps in the GIS' *relations to other variables* and recommend testing criterion-related and discriminant validity in studies of selecting/training transnational employees or global government workers. Meanwhile, GIS scores correlate moderately with cosmopolitan behaviors ($r = 0.22$–$0.39$), but that indicator merely indexed some items that Türken and Rudmin had excluded previously. Strong negative associations between GIS scores and measures of right-wing authoritarianism ($r = -0.41$) and social dominance orientation ($r = -0.41$), alongside strong positive associations with majority integration efforts ($r = 0.59$), and significant but small correlations

[2] Key words included combinations of the following truncations: cosmopolit*; international*; multicultural*; universal*; national*; and global* crossed with identity; attitudes; orientation; self; and values.

[3] Ozer and Schwartz (2016) reported reliability estimates with Ladakhi young adults ($M_{age} = 24.26$) in India ($n = 186$) resembled those of Türken and Rudmin, as did Nwafor, Obi-Nwosu, Adesuwa, and Okoye (2016) with Igbo undergraduates in Nigeria ($n = 300$). Thier et al. (2016) reported low alphas (0.48–0.68) with samples of U.S. high school students ($n = 53$) and International Baccalaureate Grade 11 and 12 students from 24 countries ($n = 121$). Thier et al.'s samples' scores produced alphas of > 0.78 for two other global citizenship measures: the Global Citizenship Scale (Morais & Ogden, 2011) and the Global Citizen Scale (Reysen & Katzarska-Miller, 2013). Relatedly, Nwafor et al. found adequate fit with its confirmatory factor analysis. Both Türken and Rudmin and Ozer and Schwartz found some items to load poorly.

[4] Among Türken and Rudmin's (2013) criteria for item inclusion were low rates of respondents' omissions or social desirability correlations; high standard deviations, item-total correlations, or indices of multicultural experiences and cosmopolitan behaviors; and limited words or characters.

with number of languages spoken ($r = 0.13–0.26$), suggest concurrent validity. But the lack of significant correlations between multicultural index and GIS scores for Turkish and U.S. students leaves room for doubt.

### 3.3. Situational-judgment test: Creativity thinking

Due to limited public data on the Social Games and VCAT, our evaluative process drew on peer-reviewed research where possible, but then relied on descriptions of these SJTs in conference papers and other documents, pressing our evaluative framework against the measures to examine their potentially unstated assumptions. In Table 3, we reported strengths, limitations, and tradeoffs for both SJTs under consideration.

#### 3.3.1. Measurement design

As an idea-generation assessment, Social Games builds off the tradition in divergent-thinking tests, attempting to draw out and evaluate the flexibility, fluency, originality, and elaboration of respondents' ideas compared to the sample. As such, *efficiency tradeoff* considerations for the Social Games include the number of items and respondents needed to produce a viable normative sample for scoring responses objectively. Available validated computerized semantic networks of lexical associations, such as WordNet, Word Associations Network, and IdeaFisher, provide reliable measures of associative distance of ideas—an attribute of originality—that are generated by a sample in response to the same prompt (Runco & Acar, 2010). A new promising *measure of internal consistency* applies predetermined conventional associations for specific divergent-thinking tasks (e.g., the word spoon in the Many Uses Game) to measure unconventional approaches with a high degree of objectivity (Runco & Acar, 2010), seemingly a plausible approach to scale Social Games responses, as well.

To improve the relevance of the Social Games exercise for different contexts and populations, *preparatory qualitative methods* could explore the most relevant social situations and "blunt expressions" to use in item development, improving connections for K-12 students. Another design approach that could alleviate some of the potential subjectivity in scoring is a similar creativity SJT Runco developed with a Likert-type scale to capture responses (personal communication): the "Alternative Movie Titles" test. This measure contains a scenario at two locations: a party and a high-stakes employment situation. The scenario asks respondents to rate alternative titles to a popular movie suggested by a friend (at the party) and a boss (at the job). This approach aims to detect how well an individual can identify when creative insights and possibilities are appropriate within social norms. Thus, responses should diverge on the two forms. This assessment style needs further testing to understand design merits and limitations that might inform an alternative format for Social Games.

**Table 3**
Strengths, limitations, & tradeoffs of situational-judgment tests of creative thinking and global citizenship.

| Measure | Strengths | Limitations | Tradeoffs |
|---|---|---|---|
| Social Games (2011) | Potentially low level of self-presentation, reference, and stereotype biases; potentially strong normative scoring system; real-life application and relevance for creativity | No empirical psychometrics available to date; potential experiential bias; validity threats; multicollinearity with personality traits related to situational judgment | Generalizability across different cultural and generational contexts vs. specified item design; cognitive load vs. sufficient number of items; closed-response vs. open-response items |
| Virtual Cultural Awareness Trainer (2011) | Respondent may be able to moderate cognitive load and make tasks personally meaningful; attenuates potential for stereotype threat and response style | No peer-reviewed empirical evidence; challenge of testing typical validity-reliability studies; potential for egocentric and self-presentation biases | Unconventional assessment techniques; design decisions needed to situate as formative or summative assessment; balance of respondent engagement and construct validity |

#### 3.3.2. Potential response bias

When evaluating Social Games, *self-presentation* does not emerge as a likely threat. The title and instructions encourage respondents to approach the task as an enjoyable challenge ("This is not a test; it is a game. The goal is to list as many different ways as you can for conveying the target idea."). Rather than choosing from options of appropriate responses, the Social Games challenges respondents to generate as many ways to convey a socially inappropriate blunt message (e.g., "you have body odor") in a socially acceptable way (i.e., "do you smell something?" or "have you been working out?"). Whereas an individual might choose a more socially desirable option with closed responses, an open-response prompt is less likely to draw out self-presentation bias. However, prompts are socially sensitive, so minor threats might remain (Kopcha & Sullivan, 2007).

*Egocentric bias* might be less of a threat in this SJT than in other formats that force respondents to make explicit interpersonal comparisons. Yet, one bias our evaluative framework does not explicitly include might be more salient for a Social Games respondent: *experiential bias*. We use this term to capture the degree to which respondents' scores depend upon prior experiences or background knowledge not explicitly defined within a construct. For instance, Benedek, Könen, and Neubauer (2012) found associative thinking—a theoretically critical step in original ideas—to explain half of the variance in divergent-thinking scores. The degree to which associative distance in thinking depends upon richness and depth of prior experience remains undetermined but warrants consideration. A primary difference between divergent- and convergent-thinking tests is that the former seeks original ideas and connections. The latter seeks primarily to elicit memories and experiences (Runco et al., 2014).

Still, experiential bias has been a concern with tests such as the Social Games, even if the measured construct does not explicitly require experience in a certain domain or setting. Runco and Acar (2010) found that personal experience biased fluency and originality scores on Many Uses tests that asked respondents to list as many uses as they can think of for an common object (e.g., spoon). If these intentionally decontextualized tasks revealed experience bias, it might be magnified in context-specific or value-laden tasks. As such, less socially inclined (i.e., more introverted) individuals may have less experience navigating nuanced social situations and might struggle to fit the task within their proximal contexts.

Of similar concern, certain Social Games situations might surface emotionally disturbing experiences (e.g., bullying), triggering a *stereotype threat* that could cloud out imaginative potential and intrinsic motivation. *Response-style bias* could emerge through such a task, as well. Relevance of blunt responses and social situations are likely to be higher for some, which could result in other respondents' premature closure to continued idea generation, yielding lower fluency and originality scores. Response-style bias of different kinds could emerge out of frustration, boredom, or disinterest in divergent thinking SJT tasks if they are not relevant. Csikszentmihalyi (1996) wrote about flow in which creative ideas and acts emerge organically from the full immersion of individuals in their work. Achieving flow

on demand in a divergent-thinking task might be less accessible for some. A potential strength of the Social Games, the opportunity to respond to different types of situations, might allow more individuals to find task relevance and connections. An alternate format could ask respondents to choose relevant scenario(s) and rank response options, a typical format for SJTs (Roberts et al., 2015).

*3.3.3. Validity*

Several potential studies could deepen understanding of what construct the Social Games taps. Concurrent validity with the Big Five personality inventory would investigate the shared variance between highly original and appropriate ideational behavior and dispositions that might associate with socially dependent creative thinking (e.g., extroversion and conscientiousness). A multimethod-multitrait approach that included Social Games could specify the associations between ideational behaviors and related characteristics (e.g., critical thinking, empathy, humor, and moral grounding). Given the context-dependent nature of the situations, the Social Games would require intentional adaptation to generalize cross-culturally. Testing item types among diverse populations would be essential.

### *3.4. Situational-judgment test: Global citizenship*

*3.4.1. Measurement design*

In their white paper, Soland et al. (2013) endorse Alelo products, including the VCAT, to assess interpersonal and intrapersonal skills. Given that VCATs do not include a typical test, and respondents receive continuous feedback to improve while being assessed, Soland et al. discuss how the nature of VCAT pushes the bounds of what we typically consider as assessment. Notably, VCAT's designers aim to train military personnel to negotiate foreign linguistic and cultural settings.

By incorporating ethnographic interviews and other "anthropological and linguistic research" (p. 3) and "best-practice…methods" (p. 4), Johnson et al. (2011) used *preparatory qualitative measures* throughout VCAT's development. By contrast, its *measures of internal consistency* remain unknown. Though limited in its descriptiveness, Johnson and Zaker (2012) conducted a study to learn about the virtual coaching approach that guides users through simulations and provides feedback when users make linguistic, social, or cultural errors. Though the assessment becomes a learning experience for avoiding future mistakes, this feature threatens comparability of results and traditional measures of reliability (e.g., test-retest). Alelo considered *efficiency tradeoff* in their assessment design, seeming to balance intentions to both assess the respondent without overloading the cognitive demand of each task. VCAT emphasizes immediate feedback to teach respondents along the way. When respondents err, simulated coaches, who are culturally and linguistically native to designed settings, target mistakes through body language and/or verbal feedback. These nuances may enhance relevance and both cognitive and emotional engagement of realistic scenarios. Blending curriculum with assessment, they may create more overt feedback than would be typical in K-12 classrooms.

*3.4.2. Potential response bias*

Though Alelo products are susceptible to *self-presentation* and *egocentric* biases, *stereotype threats* and *response-style* bias appear minimal hindrances. For instance, pre-assessment questionnaires determine the specific modules assigned to VCAT users. Savvy users seeking the easiest ways to achieve desirable scores or the quickest paths to completion could manipulate the VCAT by sabotaging pre-assessment placements. Furthermore, the real-time feedback mechanism can be disabled, leaving users with only summative coaching. If VCAT aims to rate learner adaptability in the face of feedback about unfamiliar cultural norms, eliminating formative coaching would remove this opportunity. As a particular strength of SJTs (see Lagattuta et al., 2012), evaluation of either behavioral tendencies or knowledge acquisition through respondents' judgments may be lost without documentation of reactions to formative coaching.

VCAT's main objective is for users to navigate realistic threats of cultural stereotypes. The simulation signals when virtual characters' attitudes change during interactions; thus, VCAT supports the development of awareness and appropriate behavioral adjustment to potentially harmful stereotypes (e.g., greetings to an old man versus a little girl; Johnson & Zaker, 2012). Alelo's coaching and intervention design signals learners' emotional engagement, further reducing the likelihood of self-presentation threats. Compared to performance-based assessment in authentic classroom contexts, the private nature of the VCAT might eliminate the potential for stereotype threat and social desirability bias. Notably, evidence from classrooms in developed nations identified positive and negative stereotypes affecting student performance on several skill assessments (Aronson & Dee, 2012). Lacking repetitive response formats (e.g., multiple choice) or redundant situation types, the VCAT may demonstrate resistance to response-style biases, as well. The VCAT allows respondents to demonstrate disparate cultural skills relevant to an array of contexts—from patrolling foreign checkpoints to coordinating humanitarian aid with multicultural officials (Johnson et al., 2011).

*3.4.3. Validity*

Soland et al. (2013) frame Alelo products within a multiple-measures format targeting overlapping skills and dispositions. Though relevant to instruction, this overlap creates critical challenges to deriving evidence that supports VCAT's *internal structures*. For instance, if respondents receive poor ratings, how can they be judged precisely on the many global citizenship-related competencies required to understand the task, behave appropriately, and adapt behavior based on immediate feedback? Are ratings poor due to lack of self-monitoring, inability to avoid linguistic faux pas, or disinterest in the cultural norms of others? Though consequences from intended and unintended test score uses and interpretations within the K-12 field remains unknown (Soland et al., 2013), Alelo's simulations may demonstrate some evidence supporting its *relations to other variables* by presenting transdisciplinary lessons that develop and evaluate critical skills and dispositions for K-12 learners in the 21st century.

Given that Alelo opposes "culture-specific training" (Johnson et al., 2011, p. 3), *consequences of test score use* remain uncertain. On one hand, Alelo's aim to develop and assess culture-general skills and attitudes can be seen as an asset for generalizability. On the other hand, by avoiding to differentiate the components of global citizenship into culture-specific bins, Alelo's approach contrasts with the idea that global citizenship learning and demonstration is "complex and occurs in a wide variety of formal, non-formal and informal learning settings" (Eidoo et al., 2011, p. 59). For instance, Alelo appears to teach and assess perspective-taking and rapport-building as general skills that transfer and apply equally across cultural contexts. In sampling active-duty and former military personnel, occupational experts, and native speakers from target geographical areas, future assessment innovators should investigate if these sampling decisions support robust concurrent and predictive validity or face validity only. At this stage, available research conducted with the VCAT leaves such questions unanswered.

### 3.5. Discrete-choice experiments

To compare against our chosen self-reports and SJTs, we designed two conceptual DCE models depicted in Figs. 2 and 3 for creative thinking and global citizenship, respectively. We applied design techniques from our evaluative framework and its data sources. After describing the item construction and design briefly, we evaluate potential response biases and validity concerns for each measure. For parsimony's sake, we do not detail potential analytic strategies for the proposed DCE measure.

#### 3.5.1. Measurement design: Creativity

To explore DCEs for creative thinking, we referred to extant instruments (Kumar, Kemmler, & Holman, 1997; Runco et al., 2001; Runco, 2014) and a common theoretical foundation: the two-tiered theory of creativity that describes the creative thinking process as an amalgam of problem finding, ideation and evaluative thinking alongside motivation and knowledge (Runco & Chand, 1995). Our DCE aims to measure seven attributes of creative-thinking behavior: (a) dependence on others, (b) problem-finding habits, (c) fluency, (d) playfulness and originality, (e) creative roadblocks, (f) practice and discipline, and (g) creative process-oriented mindset. Focused on students in Grades 10–12, we capped the number of attributes at 7 and levels at 2 to be mindful of the cognitive load required of respondents.

#### 3.5.2. Measurement design: Global citizenship

Our DCE for global citizenship uses a forced-choice format to illustrate design opportunities such as emphasizing unidimensional or multidimensional measures.[5] Generally, each statement should be equally desirable so perceived desirability does not factor into participants' decisions (Brown & Maydeu-Olivares, 2012). With dyadic choice sets, respondents make singular decisions, automatically revealing a rank ordering ahead of the unselected statement. To gain as much information as possible, DCE designers can ask respondents to choose the statement that is most or least like them with groupings of three or more. Knowing which choices are most and least like respondents provides a relative middle ranking for the unselected statement, a situation called "full rank" because it provides complete information. To demonstrate the range of possibilities that accompany DCEs, we present dyadic, triadic, and quadratic approaches, also varying the expected cognitive load.

Our global citizenship DCE borrows items verbatim from the English version of the GIS (Türken & Rudmin, 2013), allowing us to skip the typical forced-choice development step of using self-report Likert statements to target low and high levels of the construct and subconstructs of interest. We patterned our DCE on the GIS because 3 of the 10 items that formed the two subscales, cultural openness and non-nationalism, present evidence of possible cross-loadings. Our DCE might present a useful alternative for measuring potentially overlapping domains, a hallmark of measures of interpersonal and intrapersonal skills (Soland et al., 2013). As seen in Fig. 3, the first two choice sets are unidirectional; the second two are multidimensional. To facilitate demonstration, we order the choices within each set in descending order of level relative to the given subconstruct of interest.

[5] Unidimensional items compare two or more statements that represent the same construct—typically the statements represent different levels of a single construct or subconstruct of interest. Multidimensional items compare two or more statements that represent different constructs or subconstructs.

#### 3.5.3. Considerations for bias and validity

Potential and realized benefits of DCEs are many. First, by controlling variation in one attribute of the construct, we can ensure it is not correlated with another. Second, scenarios or statement sets presented to respondents can parse observable from unobservable data. Decision rules about preferences or perceptions tend to correlate strongly with the former. Third, choosing scenarios carefully and describing attributes realistically can elicit choices and draw connections for respondents, modeling perceptions and preferences more comprehensively than correlational measures (Aubusson et al., 2014). By forcing dyadic choices, DCEs may reduce response biases and reveal more realistic perceptions or preferences than Likert-scaled self-reports.

With DCEs, respondents compare attribute differences based on detailed descriptions rather than ambiguously termed ratings, common to self-reports (e.g., "not at all," "somewhat," or "very much"). Therefore, DCEs can decrease respondent likelihood of *egocentric* or *reference bias* dramatically. In the forced-choice measure of global citizenship, we reduce self-presentation bias by positioning unrelated statements together and asking respondents to rank each of the statements. Consequently, balancing sample size and number of items would be important. Past research on optimal design in DCEs found statistical efficiency (e.g., more items, smaller samples) might sacrifice respondents' choice consistency, increasing error variance (Louviere et al., 2008). Assumptions inherent in each design decision would need to undergo testing for statistical feasibility alongside issues of cognitive load.

Various response-bias types could elicit concerns. The creative thinking DCE uses a secondary question to analyze the weighted contributions of each attribute to the "best fit" choice. This final question requires respondents to reconsider their choices and re-examine the nature of each attribute across options in the choice set. Similar to contraindicative items, this *measure of internal consistency* serves the dual purpose of analytical importance and reduction of response-style biases. Respondents must pay attention, weighing options carefully. Before our carefully designed DCEs could be considered as promising approaches, evidence for *internal structures* and *relations with other variables* would need to withstand the same scrutiny to which any other measure would be held.

## 4. Discussion

We developed an evaluative framework encompassing measurement design, considerations for response bias, and validity to compare three assessment approaches: self-reports, SJTs, and DCEs. Applying this framework to measures of creative thinking and global citizenship, we supported improvement in measures of interpersonal and intrapersonal skills. In Tables 4 (creative thinking) and 5 (global citizenship), we synthesize issues that arise in self-reports and evidence for the promise of SJTs and DCEs.

### 4.1. Self-report ubiquity and opportunity

Our literature search demonstrated the saturation of self-report measures across multiple fields and constructs. In seeking concrete, comparative evaluations of measurement approaches for creative thinking and global citizenship, we demonstrated strengths in self-reports' designs but found variability when considering validity or their resistance to response biases. Notable concerns for evaluated self-reports included egocentric and/or reference biases: implicit self-rating processes might produce insights about beliefs and processes, but re-

| Scenario: Your teacher just assigned your class homework that is expected by the end of the week, but it won't be graded. The assignment asks you to think of as many possible ways that you could use and/or improve a typical household object of your choosing. Read the two options for how you could go about this task and choose the one that fits your style best. Please answer the follow-up question at the end. | | |
|---|---|---|
| Attributes | Attribute characteristics for choice sets | |
| | Choice 1 *[Low ideational alternative]* | Choice 2 *[High ideational alternative]* |
| Input from others *[Dependence]* | I think I have an object in mind but I want to check with my friend to make sure it is good enough | I make my final choice for an object and then get input and feedback from others |
| Selecting object *[Problem finding]* | I think about a few objects and choose the one that seems easiest for thinking up new uses | I compare different objects to see what is most interesting to me and choose one that is challenging and unusual |
| Inspiration sources *[Fluency]* | I think about small changes in the object and wait until a new idea "hits" me | I wonder about how the everyday item can be radically changed or used in peculiar ways |
| Play and persist *[Originality]* | I look online for other uses of the object or improvements that can be made | I work to create new ideas by combining and rearranging existing ones |
| Getting stuck *[Roadblocks]* | When I hit a roadblock in coming up with new ideas, it usually means I am done | To avoid getting stuck, I keep an open mind about my ideas and take a break if I need to |
| Multiple drafts *[Practice and discipline]* | After my first draft of ideas is made, I make sure spelling is correct and the list seems long enough and I call it done | I read through my first draft of ideas and improvements, crossing out those that are boring or too conventional |
| Attitude about product *[Self-efficacy and mindset]* | If my ideas don't seem as exciting and unique compared to others, the exercise was a failure | I enjoyed the chance to think creatively even if my final ideas weren't the best |
| 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 | |

**Fig. 2.** This choice set for a discrete choice experiment of creative thinking illustrates the low and high levels for each attribute for instructional purposes, only. In practice, choice sets would hold certain attributes constant across Choices 1 and 2 in the same set to isolate the influence of each attribute. Italicized text would not be visible to respondent.

sults might not be incomparable across contexts (e.g., peer groups or schools). As a result, psychometricians might admonish educational leaders to exercise caution before using self-report data to inform high-stakes decisions.

In addition to the SJT and DCE formats, other potential solutions exit. Still underused in most settings, King, Murray, Salomon, and Tandon (2004) introduced *anchoring vignettes* to attempt to weight egocentric and self-presentation biases that likely occur systematically for many self-report measures. We believe another understudied bias, *intrapersonal harshness*, threatens estimates from self-reports of interpersonal and intrapersonal skills. Intrapersonal harshness occurs when a respondent unwittingly, but sincerely, deflates scores as a result of acquiring greater awareness of a particular skill (e.g., as I learn, I understand better how much I still have to grow). Such a response might explain some confusing findings (e.g., West et al.'s [2014] report of students attending high-performing charter schools demonstrating lower mean scores for some interpersonal and intrapersonal skills compared to lower-performing public school students). This unexpectedly iatrogenic effect could stem from students becoming more self-critical through direct training and/or teacher feedback, ostensibly scoring lower due to teacher intervention.

Multimethod approaches with multiple reporters offer an opportunity to detect, reduce, or control for response biases in self-report, though such approaches sacrifice efficiency. Furthermore, external reports of student behaviors or attitudes (i.e., from parents or teachers) give rise to questions about whose report is most valid and what type of bias (including implicit) belongs to which reporter (see Furrer & Skinner, 2003). Another promising *measure of internal consistency*, the use of discretionary time on task measures, can evaluate respondent motivation to engage and persist (Plucker et al., 2006). This approach can provide rich data about response style, relevance of task, and other interrelated constructs, such as persistence. Response rate statistics that describe mischievous responses and systematic missingness on behavioral and attitudinal self-reports have shown promise, as well. Hitt, Trivitt, and Cheng (2014) found these variables to predict long-term outcomes, such as future income, in several nationally representative datasets. In designing self-report instruments, capitalizing on these findings might decrease threats of re-

| | | |
|---|---|---|
| Starting with Set 1, please read each of the statements the choice set contains. Then, complete each sentence below the choice set by using the letter that fits each sentence best for you. | | |
| Set 1*: Unidimensional [Cultural openness; moderate cognitive load]* | | |
| Statement A: I like listening to music from different cultures. *[Higher cultural openness]* | Statement B: I enjoy learning about different cultures. *[Moderate cultural openness]* | Statement C: I don't like listening to music from different cultures. *[Lower cultural openness]* |
| Statement _____ describes me **best**.<br>Statement _____ is **the farthest** away from describing me. | | |
| Set 2*: Unidimensional [Non-nationalism; reverse-indicated; low cognitive load]* | | |
| Statement A: I feel intense pride when I think about my country. *[Higher nationalism]* | Statement B: My country is one of the best in the world. *[Lower nationalism]* | |
| Statement _____ describes me **best**. | | |
| Set 3*: Multidimensional, [Low cognitive load]* | | |
| Statement A: I identify with a world community. *[Cultural openness]* | Statement B: My own culture is the best in the whole world. *[Nationalism]* | |
| Statement _____ describes me **best**. | | |
| Set 4*: Multidimensional, [Moderate cognitive load]* | | |
| Statement A: I consider myself more as a citizen of the world than a citizen of some nation. *[Higher cultural openness]* | Statement B: I identify with a world community. *[Lower cultural openness]* | |
| Statement C: I feel most connected to members of my own country. *[Lower nationalism]* | Statement D: One should first care for his/her nation, then others. *[Higher nationalism]* | |
| Statement _____ describes me **best**.<br>Statement _____ describes me **second** best.<br>Statement _____ is **the farthest** away from describing me. | | |

**Fig. 3.** Discrete-choice experiment forced-choice sets for global citizenship with differing levels of dimensionality and cognitive load. Italicized text would not be visible to a respondent.

sponse biases and generate important variables to enable improved models.

Moreover, meta-constructs, such as creative thinking and global citizenship, encompass multi-dimensional complexity. They include several interwoven, yet theoretically and empirically distinct, constructs. Because both meta-constructs of interest in this study are value- and culture-laden with norms and expectations, validity concerns abound (Kaufman & Beghetto, 2008; Morais & Ogden, 2011). Particularly for measures of meta-constructs, replicating our outlined process may provide critical insights when selecting from extant self-reports.

### *4.2. Considerations for situational-judgment tests*

Our two SJTs follow other similar measures that have predicted interpersonal and intrapersonal skills and procedural knowledge (Lievens, Buyse, & Sackett, 2005; Lievens & Sackett, 2012). Depending on whether an SJT aims to evaluate behavioral tendencies or knowledge acquisition, item writing must intentionally attenuate self-presentation bias. Unlike behavioral tendency prompts, responses to knowledge acquisition prompts do not necessarily represent what respondents would do in real life. However, many researchers consider the ability to evaluate and determine costs and benefits of multiple response options as precursors to actual behavior (e.g., Lievens & Patterson, 2011; Motowidlo & Beier, 2010).

As a downside, SJT items may be less accessible to diverse student populations and inefficient for formative purposes. Demonstrated by the VCAT, SJTs are not bound to text. Scenario-based structures suit SJTs to different media—images, comic strips, videos, and video games. Reduced cognitive demand is the main advantage of these modes over purely text-based presentations (e.g., Chan & Schmitt, 1997). Some research on the use of comic-style SJT items demonstrates potential for engaging respondents and reducing literacy demands (e.g., Ritzmann, Kluge, & Hagemann, 2011). To build face validity from the onset, developers should involve practitioners when possible to co-construct both scenarios and scoring keys for response options. Additionally, in analyzing SJTs, Ployhart and MacKenzie (2011) ask respondents to rate all response options for a given scenario to provide more information than asking for single responses. Given SJTs' affordances (Lipnevich, MacCann, & Roberts,

**Table 4**
Strengths, limitations, and potential for approaches to measuring creative thinking.

| Criterion | Runco Ideational Behavior Scale | Social games | Discrete-choice experiments |
|---|---|---|---|
| Preparatory qualitative measures | No data | Potential | Strength |
| Measures of internal consistency | Strength | Potential | Potential |
| Efficiency tradeoffs | No data | Potential | Potential |
| Resistance to self-presentation bias | Limitation | Strength | Strength |
| Resistance to egocentric bias | Limitation | Potential | Potential |
| Resistance to stereotype threat | Potential | Potential | No data |
| Resistance to response-style bias | Limitation | Potential | Potential |
| Internal structure | Potential | Potential | Potential |
| Consequences of use | Potential | No data | No data |
| Relations to other variables | Strength | Potential | No data |

*Note*. Strengths indicate well-established dimensions for the indicated measure. Potential indicates areas that are partially established for the indicated measure. Limitations indicate weak dimensions due to poor or limited evidence. No data indicate that there is no research to evaluate the dimension.

**Table 5**
Strengths, limitations, and potential for approaches to measuring global citizenship.

| Criterion | Global identity scale | Virtual cultural awareness trainer | Discrete-choice experiment |
|---|---|---|---|
| Preparatory qualitative measures | Strength | Strength | Strength[a] |
| Measures of internal consistency | Potential | Limitation | Potential |
| Efficiency tradeoffs | Strength | Strength | Strength |
| Resistance to self-presentation bias | Limitation | Limitation | Strength |
| Resistance to egocentric bias | Limitation | Limitation | Strength |
| Resistance to stereotype threat | Limitation | Strength | No data |
| Resistance to response-style bias | Limitation | Strength | Strength |
| Internal structure | Potential | Limitation | Potential |
| Consequences of use | Potential | Strength | No data |
| Relations to other variables | Potential | Potential | No data |

*Note*. Strengths indicate well-established dimensions for the indicated measure. Potential indicates areas that are partially established for the indicated measure. Limitations indicate weak dimensions due to poor or limited evidence. No data indicate that there is no research to evaluate the dimension.

[a] Our discrete-choice experiments strength in preparatory qualitative measures depends solely on its use of items from the Global Identity Scale, which was uncommonly strong in that area.

2013), the technique may improve (a) equity for racial/ethnic minorities, (b) sensitivity to context-general and context-specific processes, (c) suitability for formative assessments, (d) robustness to some response bias issues, and (e) relevance and enjoyment for students.

### *4.3. Considerations for discrete-choice experiments*

As our two conceptual DCEs demonstrate, concerns of self-presentation bias remain; yet, careful iterations and refinements can reduce that bias (Asplund, Lopez, Hodges, & Harter, 2009; Roberts et al., 2015; Cao, 2016). We provided two ways to assess interpersonal and intrapersonal skills in K-12 settings. Though the forced-choice approach reduces cognitive load to as little as one pair of attributes per set with attributes targeting the same construct, this approach greatly increases the number of sets needed and may not attenuate some response biases. In the educational context, Dweck's Implicit Personality Theory scale, has been adapted multiple times to create simplified versions of forced-choice items that can estimate learners' motivational orientations (Beckmann, Wood, Minbashian, & Tabernero, 2012; Ziegler & Stoeger, 2010). Perhaps the most widely used forced-choice assessment, the Clifton StrengthFinder® has been used in professional settings (Asplund et al., 2009). This approach uses 177 unidimensional forced-choice sets, providing two descriptive options at opposite ends of a given attribute with a follow-up question about how much the chosen attribute fits the respondent. Evidence of reliability and validity indicates this method's potential.

Though fewer unidimensional sets may limit cognitive load (Brown & Bartram, 2009), multidimensional sets targeting different constructs, plus the use of larger groupings (e.g., triads and quadratics), provide respondents with more choices. Forcing respondents to choose between two options, both of which may not be relatable, could result in discomfort and affect future responses negatively. By increasing the number of statement options per choice set, a developer may reduce the number of forced-choice items needed in the assessment, thus assuaging test fatigue. In addition to designing the number of statements per set, developers determine the number of decisions a respondent must make about those statements. Will respondents rank each option, or just select the option closest or farthest from their typical behaviors, attitudes, preferences, or beliefs? Forced-choice items can provide more information as numbers of statements and decisions increase. For instance, pairs provide one comparison; full-rank triads provide three.

Notably, the development of forced-choice items can account for social desirability during the designing phase. Developers can administer a pool of Likert-type statements needed to build forced-choice items from a group of respondents (i.e., students) who receive "fake good" instructions, in which they are asked to respond as if they are trying to impress a teacher. Logically, this approach suggests that ratings would be higher for those statements that are more desirable in this context and lower for those that are less desirable when compared to normal conditions. By grouping statements together that experience comparable change under "fake good" conditions, forced-choice items can account for desirability.[6] Importantly, an array of design opportunities reflects the varied decisions during development and administration; many more exist across analytic approaches.

DCEs could also provide another benefit: accounting for inputs of school quality that are traditionally absent from accountability systems in education. By measuring the environmental and instructional factors that may lead to interpersonal and intrapersonal skill development, DCEs could generate important data for the next generation of accountability systems. Practitioners and researchers could assess students' perceived skill development, the contextual factors that support or stifle such self-perception, and student access to productive opportunities to learn and apply the skill(s). Given that DCEs, like SJTs, may represent real life more accurately because choices abound, these approaches are likely harder to fake and potentially more sensitive to changes in student perceptions, behaviors, and attitudes than self-reports.

[6] Türken and Rudmin (2013) note that forcing choice might compel some respondents to show "loyalty to the wider world, transcending their local and national boundaries" (p. 69). This effect would be useful for measurement, but Kunst and Sam (2013) noted that it might present an "acculturation dilemma" for ethnic minorities who are torn between heritage maintenance and global assimilation (p. 4) when presented as a forced choice.

### *4.4. Conclusion*

This study's evaluative framework explored important considerations for evaluating the measures. Given that measures should provide actionable information to educators and their students, designers and consumers of measures should strive for instruments to be grounded in authentic learning contexts. Beyond strictly research purposes, measures should inform effective teaching strategies and learning environment designs to support skill development. As the constellation of skills esteemed in accountability models broadens to include interpersonal and intrapersonal skills, developing and vetting effective measures requires a reinvigorated effort. Only through those efforts will promising innovations for education research such as SJTs and DCEs move from the fringe to the norm of practice.

## Uncited references

Johnson and Wu, 2008
The Secretary's Commission on Achieving Necessary Skills, 1991

## References

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 (5), 219–229.

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.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing, 2014. Standards for educational and psychological testing. AERA, Washington, DC.

Ames, M., Runco, M., 2005. Predicting entrepreneurship from ideation and divergent thinking. Creativity and Innovation Management 14, 311–315.

Anderson, R.C., 2015. The makers: Creativity and entrepreneurship. In: Zhao, Y. (Ed.), Counting what counts: Reframing education evaluation. Bloomington, IN, Solution Tree, pp. 93–108.

Aronson, J., Dee, T., 2012. Stereotype threat in the real world. In: Inzlicht, M., Schmader, T. (Eds.), Stereotype threat: Theory, processes, and application. Oxford, UK, Oxford, pp. 264–279.

Asplund, J., Lopez, S., Hodges, T., Harter, J., 2009. The Clifton StrengthsFinder 2.0 technical report: Development and validation. Gallup Consulting, Washington, DC.

Batey, M., Chamorro-Premuzic, T., Furnham, A., 2010. Individual differences in ideational behavior: Can the big five and psychometric intelligence predict creativity scores?. Creativity Research Journal 22 (1), 90–97.

Beckmann, N., Beckmann, J., Elliott, J., 2009. Self-confidence and performance goal orientation interactively predict performance in a reasoning test with accuracy feedback. Learning and Individual Differences 19, 277–282.

Beckmann, N., Wood, R., Minbashian, A., Tabernero, C., 2012. Small group learning: Do group members' implicit theories of ability make a difference?. Learning and Individual Differences 22, 624–631.

Beghetto, R.A., 2016. Creative learning: A fresh look. Journal of Cognitive Education and Psychology 15 (1), 6–23.

Benedek, M., Könen, T., Neubauer, A.C., 2012. Associative abilities underlying creativity. Psychology of Aesthetics, Creativity, and the Arts 6, 273–281.

Brown, A., Bartram, D., 2009. Development and psychometric properties of OPQ32r. Supplement to the OPQ32 technical manual. SHL Group, Thames Ditton, England.

Brown, A., Maydeu-Olivares, A., 2012. Fitting a Thurstonian IRT model to forced-choice data using Mplus. Behavior Research Methods 44, 1,135–1,147.

Bunnell, T., 2009. The international baccalaureate in the USA and the emerging 'culture war'. Discourse: Studies in the Cultural Politics of Education 30 (1), 61–72.

Cao, M., 2016. Examining the fakability of forced-choice individual differences measures (Doctoral dissertation). University of Illinois at Urbana-Champaign.

Chan, D., Schmitt, N., 1997. Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. Journal of Applied Psychology 82, 143–159.

Conley, D.T., Darling-Hammond, L., 2013. Creating systems of assessment for deeper learning. Stanford Center for Opportunity Policy in Education, Stanford, CA.

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.

Csikszentmihalyi, M., 1996. Creativity: Flow and the psychology of discovery and invention. HarperCollins, New York, NY.

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, 1,123–1,138.

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.

DeVellis, R.F., 2003. Scale development: Theory and applications. Thousand Oaks, CA, Sage.

Drasgow, F., Stark, S., Chernyshenko, O.S., Nye, C.D., Hulin, C., White, L.A., 2012. Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions. U.S. Army Research Institute for the Behavioral and Social Sciences, Fort Belvoir, VA.

Duckworth, A., Yeager, D., 2015. Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. Educational Researcher 44, 237–251.

Duncan, A., 2013, November. Building a stronger pipeline of globally competent citizens. In: Speech presented at the international education week "mapping the nation: Making the case for global competency" launch event. DC, Washington.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53 (1), 109–132.

Eidoo, S., Ingram, L.A., MacDonald, A., Nabavi, M., Pashby, K., Stille, S., 2011. Through the kaleidoscope: Intersections between theoretical perspectives and classroom implications in critical global citizenship education. Canadian Journal of Education 34 (4), 59–85.

Engel, S., 2009. How teachers respond to children's inquiry. American Educational Research Journal 46, 183–202.

Every Student Succeeds Act of 2015, 20 U.S.C. §1005 (2015)

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., … Beechum, N.O., 2012. Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance: A critical literature review. University of Chicago Consortium on Chicago School Research, Chicago.

Florida, R., 2002. The rise of the creative class: And how it's transforming work, leisure, community, and everyday life. Basic, New York, NY.

Fukuda, E., Anderson, R.C., Lench, S., 2015. Understanding Maine's guiding principles. Maine Department of Education, August, ME.

Furrer, C., Skinner, E., 2003. Sense of relatedness as a factor in children's academic engagement and performance. Journal of Educational Psychology 95 (1), 148–162.

Heckman, J.J., 2000. Policies to foster human capital. Research in Economics 54 (1), 3–56.

Hitt, C., Trivitt, J., Cheng, A., 2014. When you say nothing at all: The surprisingly predictive power of student effort on surveys. (EDRE Working Paper No. 2014-10).

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.

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.

Jackson, D.N., Wroblewski, V.R., Ashton, M.C., 2000. The impact of faking on employment tests: Does forced-choice offer a solution?. Human Performance 13, 371–388.

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. Clarion Events, London, UK.

Johnson, W.L., Zaker, S.B., 2012. The power of social simulation for Chinese language teaching. In: Proceedings of TCLT7. (Honolulu, HI).

Johnson, W.L., Wu, S.M., 2008. In: Woolf, B., Aimeur, E., Nkambou, R., Lajoie, S. (Eds.), Assessing aptitude for learning with a serious game for foreign language and culture. International . conference on intelligent tutoring systems. Springer-Verlag, Berlin.

Kaufman, J., Beghetto, R., 2008. Exploring "mini-C": Creativity across cultures. In: DeHaan, R., Venkat Narayan, K. (Eds.), Education for innovation: Implications for India, China, and America. Sense, Rotterdam, The Netherlands, pp. 165–180.

Kennelly, B., Flannery, D., Considine, J., Doherty, E., Hynes, S., 2014. Modeling the preferences of students for alternative assignment designs using discrete choice experiment methodology. Practical Assessment, Research & Evaluation 19 (16).

Killick, D., 2011. Seeing ourselves-in-the-world: Developing global citizenship through international mobility and campus community. Journal of Studies in International Education **16**, 2011, 372–389 (1028315311431893).

Kim, K., 2011. The creativity crisis: The decrease in creative thinking scores on the Torrance Tests of Creative Thinking. Creativity Research Journal 23, 285–295.

King, G., Murray, C., Salomon, J., Tandon, A., 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. American Political Science Review 98 (1), 191–207.

Kopcha, T., Sullivan, H., 2007. Self-presentation bias in surveys of teachers' educational technology practices. Educational Technology Research & Development 55, 627–646.
Kumar, V.K., Kemmler, D., Holman, E.R., 1997. The creativity styles questionnaire—Revised. Creativity Research Journal 10 (1), 51–58.
Kunst, J.R., Sam, D.L., 2013. Expanding the margins of identity: A critique of marginalization in a globalized world. International Perspectives in Psychology: Research, Practice, Consultation 2 (4), 1–17.
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.
Lau, S., Roeser, R.W., 2008. Cognitive abilities and motivational processes in science achievement and engagement: A person-centered analysis. Learning and Individual Differences 18, 497–504.
Lievens, F., Buyse, T., Sackett, P.R., 2005. The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. Journal of Applied Psychology 90, 442–452. http://dx.doi.org/10.1037/0021-9010.90.3.442.
Lievens, G., Patterson, F., 2011. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high stakes selection. Journal of Applied Psychology 96, 927–940.
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.
Lipnevich, A.A., MacCann, C., Roberts, R.D., 2013. Assessing noncognitive constructs in education: A review of traditional and innovative approaches. In: Saklofske, D.H., Reynolds, C.B., Schwean, V.L. (Eds.), Oxford handbook of child psychological assessment. Oxford University, Cambridge, UK, pp. 750–772.
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. 35, 360–375.
Maine Department of Education, 2015, September 9. Maine DOE continues efforts with guiding principles. Retrieved from http://mainedoenews.net/2015/09/09/maine-doe-continues-efforts-with-guiding-principles/ (Web log post).
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.
Motowidlo, S.J., Beier, M.E., 2010. Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. Journal of Applied Psychology 95, 321–333.
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 Pellegrino, J.W., Hilton, M.L. (Eds.), Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. National Academies, Washington, D.C..
Nett, U.E., Goetz, T., Daniels, L.M., 2010. What to do when feeling bored? Students' strategies for coping with boredom. Learning and Individual Differences 20 (6), 626–638.
Nwafor, C.E., Obi-Nwosu, H., Adesuwa, A., Okoye, C.A., 2016. Toward globalization: Construct validation of Global Identity Scale in a Nigerian sample. Psychology & Society 8 (1), 85–99.
Osterlind, S.J., 2009. Theory, principles, and applications of mental appraisal. Pearson, Boston, MA.
Ozer, S., Schwartz, S.J., 2016. Measuring globalization-based acculturation in Ladakh: Investigating possible advantages of a tridimensional acculturation scale. International Journal of Intercultural Relations 53, 1–15.
Pannells, T., Claxton, A., 2008. Happiness, creative ideation, and locus of control. Creativity Research Journal 20 (1), 67–71.
Pecheone, R., Kahl, S., Hamma, J., Jaquith, A., 2010. Through a looking glass: Lessons learned and future directions for performance assessment. Stanford Center for Opportunity Policy in Education, Stanford, CA.
Perna, L.W., May, H., Yee, A., Ransom, T., Rodriguez, A., Fester, R., 2013. Unequal access to rigorous high school curricula: An exploration of the opportunity to benefit from the international baccalaureate diploma programme. Educational Policy 29, 402–425.
Provasnik, S., KewalRamani, A., Coleman, M.M., Gilbertson, L., Herring, W., Xie, Q., 2007. Status of education in rural America (NCES 2007-040). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.
Ployhart, R.E., MacKenzie, W.J., 2011. Situational judgment tests: A critical review and agenda for the future. In: Zedeck, S. (Ed.), APA handbook of industrial organizational psychology: Selecting and developing members for the organization. American Psychological Association, Washington, DC, pp. 237–252.
Plucker, J., Runco, M., Lim, W., 2006. Predicting ideational behavior from divergent thinking and discretionary time on task. Creativity Research Journal 18 (1), 55–63.
Reimers, F., 2009. Global competency: Is imperative for global success. The Chronicle of Higher Education 55 (21), A29.
Reysen, S., Katzarska-Miller, I., 2013. A model of global citizenship: Antecedents and outcomes. International Journal of Psychology 48, 858–870.
Ritzmann, S., Kluge, A., Hagemann, V., 2011. Using comics as a transfer support tool for crew resource management training. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 55 (2), 118–122. http://dx.doi.org/10.1177/1071181311551442.
Roberts, R., Martin, J., Olaru, G., 2015. A Rosetta Stone for noncognitive skills: Understanding, assessing, and enhancing noncognitive skills in primary and secondary education. Asia Society & Professional Examination Service, New York, NY.
Rothstein, R., 2004. Accountability for noncognitive skills: Society values traits not covered on academic tests, so why aren't they measured in school?. The School Administrator 61 (11), 29–33.
Runco, M., 1986. Divergent thinking and creative performance in gifted and nongifted children. Educational and Psychological Measurement 46, 375–384.
Runco, M., Acar, S., 2010. Do tests of divergent thinking have an experiential bias. Psychology of Aesthetics, Creativity, and the Arts 4 (3), 144–148.
Runco, M., Plucker, J., Lim, L., 2001. Development and psychometric integrity of a measure of ideational behavior. Creativity Research Journal 13, 393–400.
Runco, M., Walczyk, J., Acar, S., Cowger, E., Simundson, M., Tripp, S., 2014. The incremental validity of a short form of the ideational behavior scale and usefulness of distractor, contraindicative, and lie scales. Journal of Creative Behaviour 48, 185–197.
Runco, M., 2014. Runco creativity assessment battery: Social games assessment. Creativity Testing Services, Athens, GA.
Runco, M.A., Chand, I., 1995. Cognition and creativity. Educational Psychology Review 7, 243–267.
Runco, M.A., Jaeger, G.J., 2012. The standard definition of creativity. Creativity Research Journal 24 (1), 92–96.
Sheehy-Skeffington, J., Capturing sociological concepts with psychological rigor: a commentary on Türken & Rudmin (2013). *Psychology & Society*, **5**(2), 2013, 90–93.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. Asia Society & RAND, New York, NY.
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.
Sternberg, R., 2006. The nature of creativity. Creativity Research Journal 18 (1), 87–98. http://dx.doi.org/10.1207/s15326934crj1801_10.
Strahan, R., Gerbasi, K.C., 1972. Short, homogeneous versions of the Malowe-Crowne social desirability scale. Journal of Clinical Psychology 28 (2), 191–193.
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 and Developmental Disorders 44, 2,375–2,380.
The Secretary's Commission on Achieving Necessary Skills, 1991. What work requires of schools: A SCANS report for America 2000. U.S. Department of Labor, Washington, DC.
Thier, M., 2016. Left behind: Associating school-level variables with opportunities for global education. Presented at the 2015 annual conference of the Australian Association for Research in Education (AARE); posted in 2016 http://www.aare.edu.au/publications-database.php/9769/left-behind-associating-school-level-variables-with-opportunities-for-global-education.
Thier, M., Thomas, R., Tanaka, J., Minami, L.A., 2016. Global by design: A participatory evaluation of a global citizenship after-school program. Journal of Research in Curriculum and Instruction 20 (3), 220–231.
Türken, S., Rudmin, F.W., 2013. On psychological effects of globalization: Development of a scale of global identity. Psychology and Society 5 (2), 63–89.
Wagner, T., 2012. Creating innovators: The making of young people who will change the world. Scribner, New York, NY.
Wallach, M.A., Kogan, N., 1965. Modes of thinking in young children. Holt, Rinehart & Winston, New York, NY.
Waschbusch, D.A., Cunningham, C.E., Pelham Jr., W.E., Rimas, H.L., Greiner, A.R., … 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.
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.
West, M.R., Kraft, M.A., Finn, A.S., Martin, R., Duckworth, A.L., … Gabrieli, J.D., 2014. Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. In: CESifo area conference on economics of education. CESifo, Munich.
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. Corwin, Thousand Oaks, CA.
Ziegler, A., Stoeger, H., 2010. Research on a modified framework of implicit personality theories. Learning and Individual Differences 20, 318–326.