Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes

Angela L. Duckworth[1] and David Scott Yeager[2]

There has been perennial interest in personal qualities other than cognitive ability that determine success, including self-control, grit, growth mind-set, and many others. Attempts to measure such qualities for the purposes of educational policy and practice, however, are more recent. In this article, we identify serious challenges to doing so. We first address confusion over terminology, including the descriptor *noncognitive*. We conclude that debate over the optimal name for this broad category of personal qualities obscures substantial agreement about the specific attributes worth measuring. Next, we discuss advantages and limitations of different measures. In particular, we compare self-report questionnaires, teacher-report questionnaires, and performance tasks, using self-control as an illustrative case study to make the general point that each approach is imperfect in its own way. Finally, we discuss how each measure's imperfections can affect its suitability for program evaluation, accountability, individual diagnosis, and practice improvement. For example, we do not believe any available measure is suitable for between-school accountability judgments. In addition to urging caution among policymakers and practitioners, we highlight medium-term innovations that may make measures of these personal qualities more suitable for educational purposes.

Keywords: accountability; assessment; character; improvement research; measurements; noncognitive; psychological assessment; psychology

Measurement matters. Although reason and imagination also advance knowledge (Kuhn, 1961), only measurement makes it possible to observe patterns and to experiment—to put one's guesses about what is and is not true to the test (Kelvin, 1883). From a practical standpoint, intentionally changing something is dramatically easier when one can quantify with precision how much or how little of it there is (Drucker, 1974).

In recent years, scholars, practitioners, and the lay public have grown increasingly interested in measuring and changing attributes other than cognitive ability (Heckman & Kautz, 2014a; Levin, 2013; Naemi, Burrus, Kyllonen, & Roberts, 2012; Stecher & Hamilton, 2014; Tough, 2013; Willingham, 1985). These so-called noncognitive qualities are diverse and collectively facilitate goal-directed effort (e.g., grit, self-control, growth mind-set), healthy social relationships (e.g., gratitude, emotional intelligence, social belonging), and sound judgment and decision making (e.g., curiosity, open-mindedness). Longitudinal research has confirmed such qualities powerfully predict academic, economic, social, psychological, and physical well-being (Almlund, Duckworth, Heckman, & Kautz, 2011; Borghans, Duckworth, Heckman, & ter Weel, 2008; Farrington et al., 2012; J. Jackson, Connolly, Garrison, Levin, & Connolly, 2015; Moffitt et al., 2011; Naemi et al., 2012; Yeager & Walton, 2011).

We share this more expansive view of student competence and well-being, but we also believe that enthusiasm for these factors should be tempered with appreciation for the many limitations of currently available measures. In this essay, our claim is not that everything that counts can be counted or that everything that can be counted counts. Rather, we argue that the field urgently requires much greater clarity about how well, at present, it is able to count some of the things that count.

[1]University of Pennsylvania, Philadelphia, PA
[2]University of Texas at Austin, Austin, TX

Educational Researcher, Vol. 44 No. 4, pp. 237–251
DOI: 10.3102/0013189X15584327

A Rose by Any Other Name: Naming and Defining the Category

Reliable and predictive performance tasks to assess academic aptitude (i.e., the capacity to acquire new academic skills and knowledge) and academic achievement (i.e., previously acquired skills and knowledge) have been available for well over a century (R. Roberts, Markham, Matthews, & Zeidner, 2005). The influence of such measures on contemporary educational policy and practice is hard to overstate.

Yet parallel measures for human attributes other than cognitive ability have not followed suit. Notably, pioneers in the measurement of cognitive ability shared the intuition that these other qualities were crucial to success both in and out of the classroom. For instance, the creators of the first valid IQ test wrote that success in school "admits of other things than intelligence; to succeed in his studies, one must have qualities which depend especially on attention, will, and character" (Binet & Simon, 1916, p. 254). The author of the widely used Weschler tests of cognitive ability likewise observed that "in addition to intellective there are also definite non-intellective factors which determine intelligent behavior" (Weschler, 1943, p. 103). Our guess is that the present asymmetry represents more of an engineering problem than a difference in importance: Attributes other than cognitive ability are just as consequential but may be harder to measure (Stecher & Hamilton, 2014).

Of the descriptor *noncognitive*, Easton (2013) has pointed out, "Everybody hates this term but everyone knows roughly what you mean when you use it." Where did the term originate? Messick (1979) explains: "Once the term *cognitive* is appropriated to refer to intellective abilities and subject-matter achievement in conventional school areas . . . the term *noncognitive* comes to the fore by default to describe everything else" (p. 282). The term is problematic. Arguably too broad to be useful, this terminology also seems to imply that there are features of human behavior that are devoid of cognition.[1] On the contrary, every facet of psychological functioning, from perception to personality, is inherently "cognitive" insofar as processing of information is involved. For example, self-control, a canonical "noncognitive" attribute, depends crucially on how temptations are represented in the mind. Cognitive strategies that recast temptations in less alluring terms (e.g., thinking about a marshmallow as a fluffy white cloud instead of a sticky, sweet treat) dramatically improve our ability to resist them (Fujita, 2011; Mischel et al., 2011). And exercising self-control also relies on executive function, a suite of top-down cognitive processes, including working memory (Blair & Raver, 2015; Diamond, 2013). Hence, from a psychological perspective, the term is simply inaccurate.

Given such obvious deficiencies, several alternatives have emerged. Without exception, these terms have both proponents and critics. For example, some prefer—whereas others, with equal fervor, detest—the terms *character* (Berkowitz, 2012; Damon, 2010; Peterson & Seligman, 2004; Tough, 2013), *character skills* (Heckman & Kautz, 2014b), or *virtue* (Kristjánsson, 2013; for a review of moral character education, see Lapsley & Yeager, 2012). To speak of character or virtue is, obviously, to speak of admirable and beneficial qualities. This usefully ties contemporary efforts toward the cultivation of such positive qualities to venerated thinkers of the past, from Plato and Aristotle to Benjamin Franklin and Horace Mann to Martin Luther King Jr., who in 1947 declared, "Intelligence plus character—that is the goal of true education."

Many educators, however, prefer terminology without moral connotations. Some have adopted the term *social and emotional learning* [SEL] *competencies*, a phrase that highlights the relevance of emotions and social relationships to any complete view of child development (Durlak, Domitrovich, Weissberg, & Gullotta, 2015; Elias, 1997; Weissberg & Cascarino, 2013). SEL terminology has grown increasingly popular in education, and a search on Google Ngram shows that mention of the phrase *social and emotional learning* has increased 19-fold in published books since its introduction in 1994 (Merrell & Gueldner, 2010). The SEL moniker may, however, inadvertently suggest a distinction from academic priorities, even though the data show that children perform better in school when SEL competencies are developed (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011).

Psychologists who study individual differences among children might alternatively suggest the terms *personality, dispositions*, and *temperament*. But such "trait" terminology may incorrectly suggest that these attributes cannot be changed by people's experiences, and the connotation of immutability is at odds with both empirical evidence (Caspi, Roberts, & Shiner, 2005; B. Roberts & DelVecchio, 2000; B. Roberts, Walton, & Viechtbauer, 2006) and pedagogical aims (Tough, 2011). Indeed, widespread interest in personal qualities is fueled in large part by the assumption that students can learn, practice, and improve them.[2]

Next, the terms *21st century skills, 21st century competencies*, and *new basic skills* have made their timely appearance (Murnane & Levy, 1996; Pellegrino & Hilton, 2012; Soland, Hamilton, & Stecher, 2013). Likewise, some authors have used the terms *soft skills* (Heckman & Kautz, 2012). Unlike trait terminology, "skill" terminology usefully connotes malleability. However, referring to skills may implicitly exclude beliefs (e.g., growth mind-set), values (e.g., prosocial motivation), and other relational attitudes (e.g., trust). The narrowness of skill terminology is obvious when considering attributes like gratitude, generosity, and honesty. Yes, these behaviors can be practiced and improved, but an authentic desire to be grateful, generous, and/or honest is an essential aspect of these dispositions. As far as the descriptor *21st century* or *new* is concerned, it seems fair to question whether attributes like self-control and gratitude—of central concern to every major philosophical and religious tradition since ancient times—are of special relevance to modernity. Indeed, these may be more timeless than timely.

Finally, all of these terms—virtues, traits, competencies, or skills—have the disadvantage of implying that they are consistently demonstrated across all possible life situations. But they are not (Fleeson & Noftle, 2008; Mischel, 1968; Ross, Lepper, & Ward, 2010; Ross & Nisbett, 1991; Wagerman & Funder, 2009). For instance, self-control is undermined when people are laboring under the burden of a negative stereotype (Inzlicht & Kang, 2010) or when authority figures are perceived as unreliable (Kidd, Palmeri, & Aslin, 2013; Mischel, 1961). Learners are grittier when they have been asked to reflect on their purpose in life (Yeager et al., 2014), and organizations can create a fixed

Table 1
Serious Limitations of Questionnaires and Performance Tasks

Serious limitations of self-report and teacher report questionnaires
Misinterpretation by participant: Student or teacher may read or interpret the item in a way that differs from researcher intent
Lack of insight or information: Student or teacher may not be astute or accurate reporters of behaviors or internal states (e.g., emotions, motivation) for a variety of reasons
Insensitivity to short-term changes: Questionnaire scores may not reflect subtle changes over short periods of time
Reference bias: The frame of reference (i.e., implicit standards) used when making judgments may differ across students or teachers
Faking and social desirability bias: Students or teachers may provide answers that are desirable but not accurate
Serious limitations of performance tasks
Misinterpretation by researcher: Researchers may make inaccurate assumptions about underlying reasons for student behavior
Insensitivity to typical behavior: Tasks that optimize motivation to perform well (i.e., elicit maximal performance) may not reflect behavior in everyday situations
Task impurity: Task performance may be influenced by irrelevant competencies (e.g., hand–eye coordination)
Artificial situations: Performance tasks may foist students into situations (e.g., doing academic work with distracting video games in view) that they might proactively avoid in real life
Practice effects: Scores on sequential administrations may be less accurate (e.g., because of increased familiarity with task or boredom)
Extraneous situational influences: Task performance may be influenced by aspects of environment in which task is performed or by physiological state (e.g., time of day, noise in classroom, hunger, fatigue)
Random error: Scores may be influenced by purely random error (e.g., respondent randomly marking the wrong answer)

mind-set climate that undermines employee motivation independently of employees' own prior mind-set beliefs (Murphy & Dweck, 2009).

We believe that all of the above terms refer to the same conceptual space, even if connotations (e.g., morality, mutability, or consistency across settings) differ. Crucially, all of the attributes of interest are (a) conceptually independent from cognitive ability, (b) generally accepted as beneficial to the student and to others in society, (c) relatively rank-order stable over time in the absence of exogenous forces (e.g., intentional intervention, life events, changes in social roles), (d) potentially responsive to intervention, and (e) dependent on situational factors for their expression.

From a scientific perspective, agreement about the optimal terminology for the overarching category of interest may be less important than consensus about the specific attributes in question and, in particular, their definition and measurement. Of course, a community of practice (e.g., a school district, a reform movement, a networked improvement community) benefits from consensual terminology (Bryk, Gomez, Grunow, & LeMahieu, 2015). Marching under the same flag, rather than several different ones, would make more obvious the fact that many researchers and educators are working to measure and improve the same student attributes (Bryk et al., 2015; Langley et al., 2009). However, because each community of practice has its own previously established concerns and priorities, the choice of a motivating umbrella term is perhaps best left to these groups themselves and not to theoretical psychologists.

Our view is pragmatic, not ideological. We suggest that the potentially interminable debate about what to call this category of student attributes draws attention away from the very urgent question of how to measure them. In this review, we refer to *personal qualities* as shorthand for "positive personal qualities other than cognitive ability that lead to student success" (see Willingham, 1985). Of course, this terminology is provisional because it, too, has flaws. For instance, attitudes and beliefs are not quite satisfyingly described as "qualities" per se. In any case, we expect that communities of research or practice will adopt more descriptive terms as they see fit.

Advantages and Limitations of Common Measures

No measure is perfect. We attempt here an incomplete sketch of the limitations and advantages of three common approaches to measuring this set of personal qualities other than cognitive ability: (a) self-report questionnaires administered to students, (b) questionnaires administered to teachers about their students, and (c) performance tasks. Throughout, we illustrate our points with one important and well-studied personal quality—self-control. *Self-control* refers to the regulation of attention, emotion, and behavior when enduringly valued goals conflict with more immediately pleasurable temptations. This is an informative example because research on self-control is burgeoning (Carlson, Zelazo, & Faja, 2013). Moreover, longitudinal research supports earlier speculation (Freud, 1920) that self-control is essential to success in just about every arena of life, including academic achievement (de Ridder, Lensvelt-Mulders, Finkenauer, Stok, & Baumeister, 2012; Duckworth & Carlson, 2013; Mischel, 2014; Moffitt et al., 2011). Where appropriate, we draw on other examples, such as grit or growth mind-set. With these few brushstrokes, summarized in Table 1 and discussed briefly below, we hope to depict the contemporary landscape of measurement as we see it.

Self-Report and Teacher-Report Questionnaires

For good reason, self-report and teacher-report questionnaires are the most common approaches to assessing personal qualities among both researchers and practitioners. Questionnaires are cheap, quick, reliable, and in many cases, remarkably predictive of objectively measured outcomes (Connelly & Ones, 2010; Duckworth, Tsukayama, & May, 2010; Hightower et al., 1986; J. Jackson et al., 2015; Lucas & Baird, 2006; B. Roberts, Kuncel,

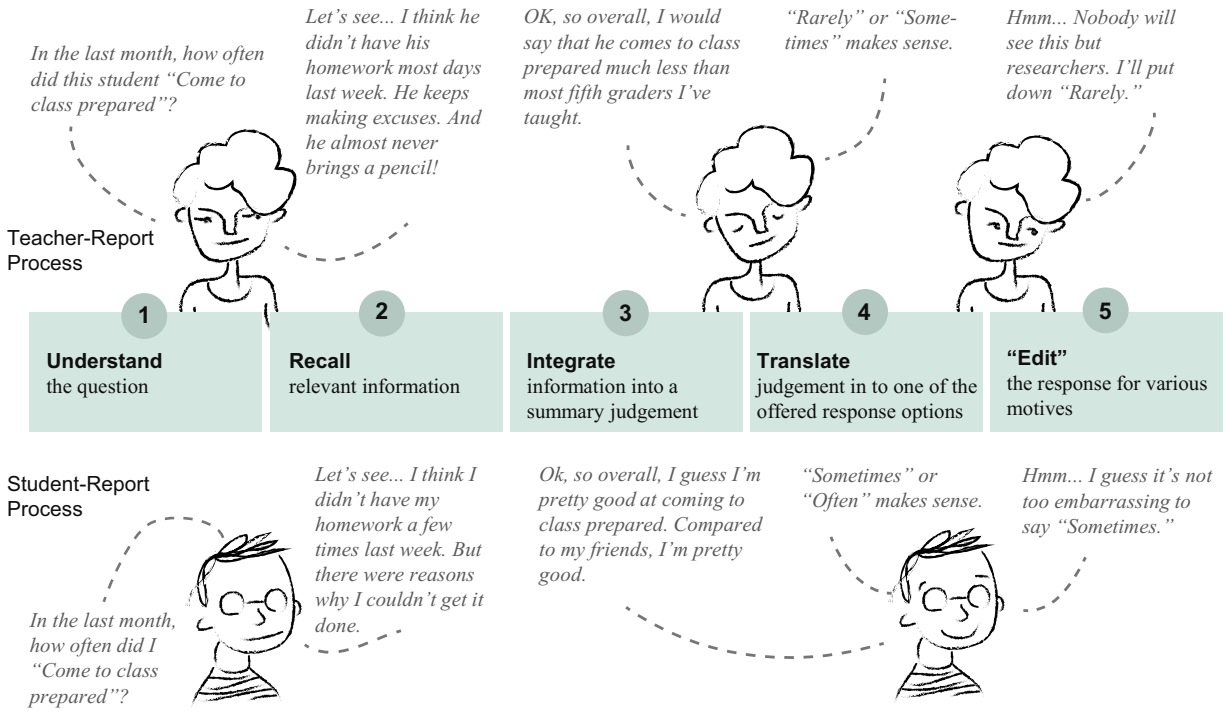

FIGURE 1. *The process by which students and teachers respond to questionnaire items.*

Shiner, Caspi, & Goldberg, 2007). Furthermore, a very large literature in social and cognitive psychology confirms that people are relatively good at using questionnaires to communicate their true opinions—provided that they in fact have answers for the questions asked and feel comfortable reporting accurately on them (see Krosnick,1999; Krosnick & Fabrigar, in press; Schuman & Presser, 1981). Indeed, self-report questionnaires are arguably better suited than any other measure for assessing internal psychological states, like feelings of belonging.

Questionnaires typically ask individuals to integrate numerous observations of thoughts, feelings, or behavior over a specified period of time ranging from "at this moment" to "in general." For example, the Character Growth Card includes a self-control item that reads, "During the past marking period, I came to class prepared" and provides response options ranging from *almost never* to *almost always* (Park, Tsukayama, Patrick, & Duckworth, 2015).

The process by which students answer this question or any other self-report item is depicted in Figure 1: (1) Students must first read and understand the question, then (2) search their memories for relevant information, (3) integrate whatever information comes to mind into a summary judgment, (4) translate this judgment into one of the offered response options, and finally, (5) "edit" their response if motivated to do so (Krosnick & Presser, 2010; Schwarz & Oyserman, 2001; Tourangeau, Rips, & Rasinski, 2000). Teacher-report questionnaires work the same way, except that it is the teacher who integrates observations of the student over time and arrives at a judgment with respect to his or her own standards. Individuals can carry out this kind of self-judgment and other-judgment arithmetic with admirable accuracy and precision (Funder, 2012).

A catalogue of threats to validity can be accomplished by considering potential failures at each stage. For (1) encoding the meaning of the questionnaire items, literacy is an obvious concern, particularly for younger or lower-achieving students. Beyond vocabulary, it cannot be assumed that students always understand the pragmatic meaning—the intended idea—of questionnaire items. For example, self-control questionnaires aim to assess the self-initiated regulation of conflicting impulses (e.g., wanting to get homework done because it is important but, at the same time, wanting to play video games because they are more fun). Yet students or teachers may instead interpret items as asking about compliance with authority (e.g., following directions simply because an adult asked).

After encoding the question itself, individuals must (2) search their memories and (3) integrate recalled information into a summary judgment. For both students and teachers, mentally integrating across past observations can reduce sensitivity to how behavior is now as compared to before (for a compelling empirical example of errors in these judgments, see Bowman, 2010). Moreover, individuals tend to see themselves and others as holding consistent beliefs and attitudes over time, and this bias toward consistency can affect what information is retrieved as well as how it is evaluated (Mischel, 1968; Nisbett & Wilson, 1977; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Ross & Nisbett, 1991; Sabini, Siepmann, & Stein, 2001).

When (3) coming to a summary judgment, teachers have the benefit of a non-egocentric perspective as well as experience with many other same-age students over the course of their careers. Nevertheless, end-of-year teacher reports may be colored by first impressions and therefore underestimate change (see moderation results in Raudenbush, 1984). In addition, many teachers see their students only in the classroom setting. Because behavior can vary across contexts (Mischel, 1968; Ross & Nisbett, 1991; Tsukayama, Duckworth, & Kim, 2013), teacher observations may not agree with those made by parents, who may see their

child in every context *except* school. Not surprisingly, correlations between separate teacher ratings of student behavior tend to be higher than between parents and teachers (Achenbach, McConaughy, & Howell, 1987).

Another limitation of teacher-report questionnaires is the potential for teachers to misinterpret student behavior. People's inferences about why others act the way they do are not always accurate (e.g., Dodge, 1980). For instance, it might seem reasonable to conclude that students who reliably complete all of their homework assignments on time are highly self-controlled. Alternatively, it is possible that some assiduous students are so intrinsically motivated to do schoolwork that they do not find alternatives, like texting and video games, at all tempting. If so, it is incorrect to infer that their conscientious academic behavior represents self-control (Duckworth & Steinberg, 2015).

Teachers' ratings of students' specific qualities can also be colored by their top-down, global evaluations. For instance, teachers may think, "This is a good kid" and then conclude, "This student must be good at delaying gratification" (see Abikoff, Courtney, Pelham, & Koplewicz, 1993; Babad, Inbar, & Rosenthal, 1982; Nisbett & Wilson, 1977).

Both students and teachers must use some frame of reference to arrive at their judgments, and the problem of "reference bias" refers to frames of reference that differ systematically across respondents (Heine, Lehman, Peng, & Greenholtz, 2002). For example, the more competent an individual is in a given domain, the more stringently he or she tends to judge himself or herself (Kruger & Dunning, 1999). Frames of reference are also influenced by the norms shared within—but not necessarily across—cultures. Thus, reference bias is most readily evidenced in paradoxical inconsistencies in cross-cultural research.

Reference bias is apparent in the PISA (Program for International Student Assessment). Within-country analyses of the PISA show the expected positive association between self-reported conscientiousness and academic performance, but between-country analyses suggest that countries with higher conscientiousness ratings actually perform worse on math and reading tests (Kyllonen & Bertling, 2013). Norms for judging behavior can also vary across schools within the same country: Students attending middle schools with higher admissions standards and test scores rate themselves lower in self-control (Goldman, 2006; M. West, personal communication, March 17, 2015). Likewise, KIPP charter school students report spending more time on homework each night than students at matched control schools, and they earn higher standardized achievement test scores—but score no higher on self-report questionnaire items such as "Went to all of your classes prepared" (Tuttle et al., 2013). Dobbie and Fryer (2013) report a similar finding for graduates of the Harlem Children's Zone charter school. There can even be reference bias among students in different grade levels within the same school. Seniors in one study rated themselves higher in grit than did juniors in the same high school, but the exact opposite pattern was obtained in performance tasks of persistence (Egalite, Mills, & Greene, 2014).[3]

In the final stages of responding to questionnaire items, individuals must (4) translate their judgment into one of the offered response options. Reference bias can be a problem here, too, insofar as what one respondent considers "rarely" may be what another respondent considers "often" (Pace & Friedlander, 1982).

Next, individuals may (5) amend their response in accordance with any of a number of motivations other than truth telling. Potentially biasing reports is "acquiescence bias," the inclination, particularly among younger students, to agree with statements regardless of their actual content (Saris, Revilla, Krosnick, & Shaeffer, 2010; Soto, John, Gosling, & Potter, 2008). Individuals may also not tell the truth simply because they would be embarrassed to admit it (Jones & Sigall, 1971).

Unfortunately, many methods thought to reduce social desirability response bias instead harm validity. For example, preemptive assurances of confidentiality can backfire if they imply that questionnaires will be about sensitive and potentially embarrassing topics (Schwarz & Oyserman, 2001). Moreover, assuring individuals of their anonymity can decrease response validity by removing accountability to be honest (Lelkes, Krosnick, Marx, Judd, & Park, 2012). And attempting to make adolescents feel comfortable reporting undesirable attitudes or behaviors by suggesting that "some people do *X*; other people do *Y*" implies to adolescents that the undesirable behavior is carried out by half of their peers, and so it artificially inflates reports of that behavior through conformity processes (Yeager & Krosnick, 2011). Unfortunately, scales purporting to measure individual differences in social desirability bias do not fulfill their promise (Uziel, 2010).

Finally, there is the problem of faking. The extent to which outright faking actually reduces the validity of questionnaires in real-world situations is hotly debated (Ziegler, MacCann, & Roberts, 2011), but the possibility of deliberately inflating or deflating scores on questionnaires is incontrovertible (Sackett, 2011).

Performance Tasks

As an alternative to asking a student or teacher to report on behavior, it is possible to observe behavior through performance tasks. A performance task is essentially a situation that has been carefully designed to elicit meaningful differences in behavior of a certain kind. Observing students in the identical contrived situation eliminates the possible confound of variation in the base rates of certain types of situations. For example, it is problematic to use "time spent doing homework" as an indicator of self-control if some students are assigned more homework than others (for example, when comparing students whose teachers or schools differ). But if all students are put in a situation where they have the same opportunity to do academic work, with the same opportunity to allocate their attention to entertaining diversions, then differences in time spent on academic work can be used to index self-control (Galla & Duckworth, 2015).

The most influential performance task in the large literature on self-control is the preschool delay-of-gratification paradigm, colloquially known as the "marshmallow test" (Mischel, 2014). At the start of the task, children are presented with a variety of treats and asked to pick their favorite. Some choose marshmallows, but others choose Oreos, chocolate candies, pretzels, and so on. Next, the less preferred treats are taken away, and the experimenter makes a smaller pile (e.g., one marshmallow) and a

larger pile (e.g., two marshmallows). The experimenter asks the child whether he or she would prefer to have the small pile right away or, alternatively, to wait for the larger pile after the experimenter comes back from doing something unrelated in the hallway. The question is not which choice the child makes—in a national study of approximately 1,000 preschool children, nearly all chose the larger, delayed treat (National Institute of Child Health and Human Development, 1999)—but rather, once the decision has been made, how long the wait to obtain the larger treat can be endured. Wait time in this standardized situation correlates positively with self-control ratings by parents and caregivers and, over a decade later, predicts higher report card grades and standardized test scores, lower self-reported reckless behavior, and healthier body weight, among other outcomes (Mischel, 2014; Tsukayama et al., 2013).

An advantage of performance tasks is that they do not rely upon the subjective judgments of students or teachers. This feature circumvents reference bias, social desirability bias, acquiescence bias, and faking. Relatedly, by assaying behavior at a moment in time, task measures could be more sensitive than questionnaires to subtle changes in behavior. Not surprisingly, several major studies examining the effects of either self-control interventions or age-related changes in self-control have used performance tasks to do so (Bierman, Nix, Greenberg, Blair, & Domitrovich, 2008; Blair & Raver, 2014; Diamond & Lee, 2011; Raver et al., 2011). Likewise, experiments that attempt to manipulate self-control in the short term commonly measure change using performance tasks rather than questionnaire measures (e.g., Baumeister, Bratslavsky, Muraven, & Tice, 1998; Hagger, Wood, Stiff, & Chatzisarantis, 2010).

Of course, the advantages of performance tasks must be considered in tandem with their limitations. As is the case with teacher-reported questionnaires, performance tasks require drawing inferences about the internal motivations, emotions, and thoughts of students. For instance, is a child who refrains from playing with toys when instructed to do so exerting autonomous self-control, or does such behavior represent compliance with adult authority (see Aronson & Carlsmith, 1963; Eisenberg et al., 2004; Mischel & Liebert, 1967)? Although the task itself is "objective," interpreting performance is nevertheless "subjective" in the sense that behavior must be interpreted by the researcher.

Relatedly, a one-time performance task may be appropriate for assessing the capacity of a student to perform a certain behavior when maximally motivated to do so but not particularly diagnostic of their everyday behavior in typical life situations (Duckworth, 2009; Sackett, 2007). For many personal qualities (e.g., generosity, kindness, honesty), what matters most is how a child usually behaves, not how he or she could behave when trying his or her hardest. In these cases, performance tasks that assess behavior under optimally motivating circumstances miss the mark. Of course, for some personal qualities, assessing capacity may be appropriate, because the construct itself specifies an ability that may or may not be expressed in daily life. For example, performance task measures of emotional intelligence appropriately assess the ability—not the propensity—to perceive, understand, and manage emotions (Brackett & Geher, 2006; Brackett & Mayer, 2003).

Another limitation of performance tasks is their sensitivity to factors irrelevant to the attribute of interest. Miyake and Friedman (2012) call this the "task-impurity problem" (p. 8) and use as an example the Stroop task of executive function. Completing the Stroop task entails looking at the names of colors printed in variously colored ink. When the name of the color is different from the ink in which it is printed (e.g., the word *green* printed in red), then naming the ink color requires executive function. But executive function is not all that is required. Quick and accurate performance also requires color processing, verbal articulation, motivation to pay attention, and so on. Task impurity is thought to be one reason why performance tasks assessing executive function are only weakly correlated with questionnaire measures of self-control (Duckworth & Kern, 2011; Sharma, Markon, & Clark, 2014).

In addition, performance tasks may thrust individuals into situations they might have avoided if left to their own devices (Diener, Larson, & Emmons, 1984). Consider, for example, children faced with the dilemma of one treat now or two treats later in the delay-of-gratification task. In the test situation, children are not allowed to get up from their chair, occupy themselves with toys or books, or cover the treats with a plate or napkin. Outside of this constrained laboratory situation, any of these tactics might be employed in order to make waiting easier. In fact, more self-controlled adults say they very deliberately avoid temptations in everyday life (Ent, Baumeister, & Tice, 2015; Imhoff, Schmidt, & Gerstenberg, 2013) and as a consequence experience fewer urges to do things they will later regret (Hofmann, Baumeister, Förster, & Vohs, 2012). Thus, performance tasks foist individuals into identical circumstances so that one may assess their ability to navigate such situations, but this comes at the expense of knowing the extent to which they might have the judgment to proactively avoid or modify situations of that kind on their own (Duckworth, Gendler, & Gross, 2014).

To some extent, all performance tasks suffer from practice effects (or test-retest effects), defined broadly as the effect of repeated exposure to the same task. This is true even for the most "pure" measures of general cognitive ability (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Reeve & Lam, 2005). Familiarity with the procedures of a task can undermine score validity when the task is intended to represent an ambiguous or novel situation (Burgess, 1997; Müller, Kerns, & Konkin, 2012). For example, a first-time experience with the delay-of-gratification task is not identical to a second-time encounter because expectations of when the experimenter will return to the room are altered (McGuire & Kable, 2013). Experience with a task may also lead to boredom or increased fluency with task procedures irrelevant to the target attribute. At present, almost nothing is known about the feasibility of developing parallel forms of performance tasks assessing personal qualities for repeated administration.

Because performance tasks are standardized situations in which to observe student behavior, they must be administered under carefully controlled conditions. For example, children in the delay-of-gratification task wait longer if they trust that the experimenter is actually going to deliver on the promise of two marshmallows later (Kidd et al., 2013). Likewise, performance on self-control tasks can suffer when performed in sequence after other effortful tasks (Hagger et al., 2010). Error increases, and precision decreases, the more these situational influences differ across students.

Moreover, situational influences on task performance that vary systematically across groups create bias and potentially misleading conclusions about group differences. For example, a task that assesses diligence on academic work cannot be properly interpreted if administered in a school setting characterized by frequent noisy intrusions (e.g., students walking in and out of the testing room) or especially crowded conditions (e.g., students sitting so closely that they are distracted by each other) (Galla et al., 2014). Although questionnaire responses, too, can be influenced by transient situational influences, these effects may be small (see Lucas & Lawless, 2013). In our experience, performance tasks are especially sensitive to differences in administration, such as time of day or presence of ambient distractions.

Even when administered under optimally controlled conditions, performance tasks generate random error—the white noise produced by stochastic influences on behavior. This is especially problematic for performance tasks because most yield a single score (e.g., in the marshmallow test, the number of seconds a child can wait). Questionnaires, in contrast, usually include several different items designed to assess the same latent construct. Using multiple items exploits the principle of aggregation, which states that uncorrelated errors across items cancel out, thus reducing noise and increasing reliability (Clark & Watson, 1995; Rushton, Brainerd, & Pressley, 1983).

An obvious solution is to create a suite of different performance tasks to assess the same construct and then to aggregate results into a composite score. There are only a handful of precedents for this multitask approach to assessing self-control (Hartshorne & May, 1929; White et al., 1994). The rarity of these studies suggests that the time, expense, and effort entailed in administering a battery of performance tasks to the same children is at present prohibitive in most applied settings. A single performance task could take as many as 20 minutes to administer by a trained experimenter; doing so several times across separate sessions (to avoid fatigue) would likely require hours and hours of testing time.

Valid for What?

As the above exposition demonstrates, perfectly unbiased, unfakeable, and error-free measures are an ideal, not a reality. Instead, researchers and practitioners have at their disposal an array of measures that have distinct advantages and limitations. Accordingly, measurement experts have emphasized that validity is not an inherent feature of a measure itself but rather a characteristic of a measure with respect to a particular end use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014). Thus, different measures, with their unique advantages and limitations, are differentially valid depending not only on their psychometric properties but also on their intended application.

One important end use is basic research, and indeed, this is the purpose for which most of the measures reviewed here were developed. Given the litany of limitations noted above, it is notable that measures of personal qualities have been shown in basic research studies to be predictive of consequential life outcomes months, years, or decades later (Almlund et al., 2011; Borghans et al., 2008; Farrington et al., 2012; Moffitt et al., 2011; Naemi et al., 2012; B. Roberts et al., 2007). Of course, these research studies have sought to reject the null hypothesis of no relation between personal qualities and later life outcomes, under testing conditions where incentives to distort responses were minimal—a very different project than the applied uses we consider in this final section.

We attempt to explain how the problems with extant measures of these personal qualities can create threats to validity for more applied uses. Four common examples are *program evaluation, accountability, individual diagnosis*, and *practice improvement*. We make specific recommendations regarding each.

Program Evaluation

Many educational programs, including charter schools, in-school programming, and after-school activities, aim to cultivate self-control, grit, emotional intelligence, and other personal qualities. Yet the above review makes it clear that in many cases, self-report questionnaires have serious limitations for such evaluations. Reference bias may even produce results opposite of the truth when evaluating within-person program effects (i.e., a change from pretest to posttest) or assessing between-program differences (i.e., mean-level differences among schools or programs), as noted above (e.g., Tuttle et al., 2013; West et al., 2015).

Teacher-report measures of personal qualities may be valid when program evaluation is occurring within schools (i.e., comparing classes in the same school, where the standard for a given characteristic is presumably held constant). However, when conducting between-school program evaluation—as is common—it seems likely that self-report and teacher-report questionnaires could be biased by a nonshared frame of reference. For example, teachers at schools with more rigorous standards of behavior may rate their students more stringently.

How then should between-school program evaluations be conducted? Performance tasks may be helpful (Blair & Diamond, 2008; Greenberg, 2010). However, they have the limitations noted above, including but not limited to dependence on carefully controlled settings for proper administration, the need to tailor the task parameters to the age group, practice effects, and respondent burden. At the same time, performance tasks have perhaps the most important quality for program evaluation: objective, quantifiable behaviors that do not suffer from reference bias over time and across sites.

A potentially solvable engineering problem, in the medium term, is to create a suite of brief, scalable, age-specific performance tasks designed for group administration. This possibility was foreseen as prohibitively expensive by pioneers in the assessment of personal qualities (Hartshorne & May, 1929), but they could not have predicted the proliferation of computers and wireless technology in schools. Imagine, for example, a set of web-based tasks of academic self-control accompanied by easy-to-follow protocols and checklists for administering them (e.g., Galla et al., 2014). Assuming that practice effects (i.e., test-retest effects) could be minimized, such task batteries might allow for meaningful, apples-to-apples comparisons across schools, among individuals within schools, or within individuals over time.

In sum, scalable batteries of performance tasks to assess various personal qualities would be of great value for program

evaluation, especially as schools and districts seek to allocate limited funds wisely.

Accountability

Reference bias in questionnaire measures has a pernicious implication for accountability. Current data and theory suggest schools that promote personal qualities most ably—and raise the standards by which students and teachers at that school make comparative judgments—may show the lowest scores and be punished, whereas schools that are least effective may receive the highest scores and be rewarded for ineffectiveness (Dobbie & Fryer, 2013; O'Brien, Yeager, Galla, D'Mello, & Duckworth, 2015; West et al., 2015). Even when accountability does not carry high stakes—for instance, when between-school measures are simply used to pair high- and low-scoring schools to learn from one another—reference bias undermines school improvement: It would lead the practices from the worst schools to be spread to the best schools. Unfortunately, our experience suggests that rewriting items does not seem to eliminate this type of reference bias.

The reference bias problem alone suggests that questionnaires, as they currently exist, should not be used for between-school accountability. Yet accountability adds at least two additional concerns. First, it is not clear that aggregated student reports can reasonably distinguish among schools throughout the majority of the distribution. Indeed, even value-added measures based on standardized achievement test scores (which do not suffer from reference bias) fail to distinguish more effective from less effective teachers outside of the very high or very low ends of the distribution (Goldhaber & Loeb, 2013; Raudenbush & Jean, 2012).

One exception may be assessing personal qualities for the purpose of comparing teachers within schools. For example, the Tripod measure allows for students' ratings of different teachers in the same school; ratings of one teacher can be "anchored" using ratings of others in the same school, and these anchored ratings have been shown to correlate with differences in value-added measures among teachers within schools (Ferguson, 2012; Ferguson & Danielson, 2014; Kane & Cantrell, 2013). These measures would be excellent for identifying "positive outlier" teachers within schools—for instance, those who reduce achievement gaps and maintain a strong sense of belonging in students—and then encouraging peer teachers in the same schools to learn from their practices. Unfortunately, these measures, like many others, are not very effective when comparing between schools. This mirrors analyses of state test scores, which have found that value-added measures are better at distinguishing among different teachers in the same schools than among different teachers in different schools (Raudenbush, 2013).

There is a second, perhaps more problematic, issue with using measures for the sake of accountability: the potential for faking or unfairly manipulating data. Campbell (1976) observed, "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p. 49). Campbell was making a very general point about how good intentions can lead to unintended perverse outcomes. But this may be especially germane to self-report and teacher-report questionnaire measures, because students can easily be taught to mark the "right answer," and teachers can likewise rate their students more favorably than they really perceive them to be. Even when faking does not occur, accountability pressures for qualities such as growth mind-set can lead to superficial parroting of growth mind-set ideas, so as to increase self-reports rather than reflect true, deep, mind-set changes in students. We should note that accountability pressures can also affect performance tasks insofar as schools could be incentivized to alter testing situations to optimize student performance (for examples from achievement tests, see commentaries by Hollingworth, Dude, & Shepherd, 2010; Ravitch, 2012).

In sum, we have a simple scientific recommendation regarding the use of currently available personal quality measures for most forms of accountability: *not yet*.

Individual Diagnosis

Schools may wish to diagnose students' personal qualities to use in tracking or remediation decisions. This type of measurement involves decisions about the resources given to a child and raises two major concerns: reliability at the level of the individual person and context dependency. First, existing questionnaire measures are likely not sufficiently reliable for making diagnoses. Indeed, even more well-established clinical protocols for assessing factors such as depression rely on extensive self-report questionnaires only as a screening tool, and these have only modest sensitivity and specificity for determining whether a person is clinically depressed (Kovacs, 1992). Such measures require clinical follow-ups. Without highly reliable, multimethod, multi-informant measurement batteries whose validity has been demonstrated for diagnosis, it will be difficult for a practitioner to justify the individual diagnosis of children's personal qualities, such as self-control, grit, or growth mind-set.

Our second concern is the context dependency of some measures. Take the example of self-control. Research finds that individuals have a harder time regulating themselves when they are under stereotype threat, but they show greater self-control when they do not feel under threat (Carr & Steele, 2009; Inzlicht & Kang, 2010; Inzlicht, McKay, & Aronson, 2006). Hence, cues that signal the potential to be stereotyped might impair a stigmatized student's self-control, inappropriately supporting conclusions about the child's ability rather than about the bias in the context. This may lead to unintended effects, such as victim blaming, rather than systemic reform.

In sum, a handful of self-report or teacher-report questions cannot (currently) diagnose an individual child's self-control, growth mind-set, grit, purpose, and so on. And even if more extensive protocols were available, it would be essential to consider the possibility of situational and group-specific biases.

Practice Improvement

The ultimate goal of much education research is to systematically improve personal qualities across contexts—that is, to promote the improvement of practice (Bryk et al., 2015). Here, too, existing measures have important limitations, but they also have great potential.

As has been argued elsewhere (Bryk et al., 2015; Langley et al., 2009; Yeager & Bryk, 2014), the improvement of practice

Table 2
Summary for Practitioners and Policymakers

Conclusions

- There is a scientific consensus in the behavioral sciences that success in school and beyond depends critically on many attributes other than cognitive ability.
- As shown in Table 1, measures created for basic theory development have various advantages and limitations.
- These limitations often undermine validity for applied uses.
- Self-report and teacher-report questionnaire measures may potentially produce the opposite finding of the truth if used for between-school or within-school over-time comparisons, as in *program evaluation* and *school accountability*.
- Existing questionnaire and performance task measures are rarely sufficiently reliable to use for *diagnosis* and may produce group biases.
- Both questionnaire and performance tasks may be useful for *practice improvement* under some circumstances.

Recommendations

- A consensual umbrella term for this heterogeneous set of competencies may be less important than clarity about the individual constructs.
- A high priority for research is to improve the suite of performance tasks available for *program evaluation* and *practice improvement*.
- A second priority is to develop novel and innovative measures, capitalizing on advances in theory and technology.

requires "practical measurement." Practical measures are not measures for theory development or accountability. Instead, they are administrable in the web of daily instruction, they can be quickly reported on and communicated to practitioners, and they have direct relation to causes of student underperformance that are the explicit target of improvement efforts. They allow people to learn rapidly from practice. This means that the measures should be brief, easily collected, and also contextually appropriate. Practical measures should be sensitive to short-term changes and provide short-term feedback on progress that has or has not been made in improving personal qualities (Bryk et al., 2015; Yeager & Bryk, 2014).

Existing questionnaires demonstrate very few of these features. First, questionnaires can be quite long (Atkins-Burnett, Fernandez, Akers, Jacobson, & Smither-Wulsin, 2012). For instance, some measures of self-efficacy—a construct that could in theory be assessed with a single item—are 60 items long (Marat, 2005). Next, questionnaire measures are rarely if ever customized for different settings, and therefore their data may not be as relevant to a given teacher working in a given school. That is, issues of language and literacy, cultural norms, or even colloquialisms could compromise a practical measure. A practical measure is useful only if it helps improve practice in a given setting, not if it has reliability and validity on average in different settings. Third, conventional questionnaire measures are often not designed to be sensitive to change over time. For example, a teacher who wants to know whether classroom activities encouraged self-control during the prior week may not learn much by asking students to repeatedly respond to very general questions, such as "People say that I have iron self-discipline." At the same time, it may be possible to write optimized questions—ones that use construct-specific verbal labels to avoid acquiescence bias, use the optimal number of response options, balance bipolar choices, and so on (Gehlbach & Brinkworth, 2011; Krosnick, 1999; Schuman & Presser, 1981; Schwarz & Oyserman, 2001)—and many fewer of them to solve this latter problem (Yeager & Bryk, 2014).

We believe performance tasks can also support practice improvement. For instance, tasks can document within-person changes over the short-term. To the extent that performance tasks can be embedded online, then they may be used to produce efficient web-based reports, facilitating teachers' improvement efforts. At the same time, as noted, performance tasks still require that procedures be optimized to reduce systematic and random error. This can make them logistically difficult to embed in the web of daily practice. Still, this may be a solvable engineering problem in the medium term.

In sum, a promising area for future research is to increase knowledge of the conditions under which questionnaires and performance tasks can support the continuous improvement of educational practice.

Final Recommendations

The major conclusions of this article are summarized in Table 2. We have argued that all measures have limitations as well as advantages. Furthermore, we have observed that the applied uses of assessments are diverse, and design features that make a measurement approach helpful for one use may render it less appropriate for another. As a consequence, it is impossible to hierarchically rank measures from best to worst in any absolute sense. Rather than seek out the "most valid measure," therefore, we advise practitioners and researchers to seek out the "most valid measure for their *intended purpose*." While doing so, policymakers and practitioners in particular should keep in mind that most existing measures were developed for basic scientific research. We urge heightened vigilance regarding the use-specific limitations of any measure, regardless of prior "evidence of validity."

Whenever possible, we recommend using a plurality of measurement approaches. Although time and money are never as ample as would be ideal, a multimethod approach to measurement can dramatically increase reliability and validity (Eid & Diener, 2006; Rushton, Brainerd, & Pressley, 1983). As just one example, Duckworth and Seligman (2005) aggregated multiple measures of self-control, including a delay-of-gratification task and self-report, teacher-report, and parent-report questionnaires, finding that a composite score for self-control in the fall predicted final report card grades better than a standard measure of cognitive ability. We also encourage further innovation in measurement development. An incomplete list of promising approaches includes opportunistically mining students' online learning behavior or written communication in real time (e.g., Twitter feeds, Kahn Academy databases) for meaningful patterns of behavior (D'Mello, Duckworth, & Dieterle, 2014; Ireland &

Pennebaker, 2010; Kern et al., 2014), the aperture method of administering random subsets of questionnaire items to respondents so as to minimize administration time while maximizing content validity (Revelle, Wilt, & Rosenthal, 2010), recording and later coding 30-second audio snippets during everyday life (Mehl, Vazire, Holleran, & Clark, 2010), presenting hypothetical situations in narrative form and asking students what they would do in that circumstance (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Ployhart & MacKenzie, 2011), asking students to make observations of their peers (Wagerman & Funder, 2007), indirectly assessing personal qualities through innovative application of factor analysis to conventionally collected data (e.g., grade point average, attendance, achievement test scores; C. Jackson, 2012; Kautz & Zanoni, 2014), and contacting students throughout the day to assess their momentary actions, thoughts, and feelings (Wong & Csikszentmihalyi, 1991; Zirkel, Garcia, & Murphy, 2015). In general, efforts to advance measurement of personal qualities would greatly benefit from cross-fertilization with similar efforts in personality psychology, industrial and organizational psychology, neuroscience, and economics (Heckman & Kautz, 2014a; Pickering & Gray, 1999; Roberts, Jackson, Duckworth, & Von Culin, 2011; Schmidt, 2013; Schmidt & Hunter, 1998).

Relatedly, it has recently been suggested that supplementing questionnaires with anchoring vignettes may help reduce reference bias (King, Murray, Salomon, & Tandon, 2004; Kyllonen & Bertling, 2013). Anchoring vignettes are brief descriptions of hypothetical persons that serve as anchors for calibrating questionnaire responses. Respondents rate each vignette and then their own behavior on the same rating scale. Adjusting scores of self-report questionnaires using anchoring vignettes has been shown to resolve paradoxical findings attributed to reference bias. However, adding vignettes to questionnaires can dramatically increase respondent burden. Moreover, at present it has been impossible to verify the extent to which vignettes fully correct for reference bias (Kyllonen & Bertling, 2013).

Finally, measuring personal qualities, although difficult, is only the first step. Scientific inquiry and organizational improvement begin with data collection, but those data must be used to inform action. Too little is known about the question of how to act on data regarding the personal qualities of students in various classrooms or schools (Bryk et al., 2015). If a classroom is low in grit, what should one do? If a student is known to have a fixed mind-set, how can one intervene without stigmatizing the child (and should one intervene at all)? How can multidimensional data on personal qualities be visualized and fed to decision makers more clearly? The wise use of data in educational practice is another topic that will be increasingly important—and likely just as fraught with difficulty—as the collection of that data (Bryk et al., 2015).

Interest in the "other" side of the report card is not at all new. What is new is the expectation that one can measure, with precision and accuracy, the many positive personal qualities other than cognitive ability that contribute to student well-being and achievement. Quantifying, even imperfectly, the extent to which young people express self-control, gratitude, purpose, growth mind-set, collaboration, emotional intelligence, and other beneficial personal qualities has dramatically advanced scientific understanding of their development, impact on life outcomes, and underlying mechanisms. It is no surprise that policymakers and practitioners have grown increasingly interested in using such measures for diverse purposes other than theory development. Given the advantages, limitations, and medium-term potential of such measures, our hope is that the broader educational community proceeds forward with both alacrity and caution, and with equal parts optimism and humility.

NOTES

This research was made possible by grants to the first author from the National Institute on Aging (Grants K01-AG033182-02 and R24-AG048081-01), the Character Lab, the Gates Foundation, the Robert Wood Johnson Foundation the Spencer Foundation, and the Templeton Foundation as well as grants to the second author from the Raikes Foundation and the William T. Grant Foundation and a fellowship from the Center for Advanced Study in the Behavioral Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

[1]Interestingly, although the notion of *cognitive skills* has garnered much more adherence than the term *noncognitive skills*, both are difficult to define with precision, often misinterpreted because of lack of consensual definitions, hard to measure without influence of the other, and representative of heterogeneous rather than homogenous categories (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; Gardner, 2004; Heckman & Kautz, 2014a; Sternberg, 2008).

[2]We hasten to point out that cognitive ability is also mutable (Nisbett, 2009; Nisbett et al., 2012).

[3]Some have argued that comparisons to peers of higher or lower achievement are not merely a source of systematic measurement error but, in addition, can lead to durable changes in self-concept, motivation, and performance (Huguet et al., 2009).

REFERENCES

Abikoff, H., Courtney, M., Pelham, W. E., Jr., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*(5), 519–533.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*(2), 213–232.

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality psychology and economics* (No. w16822). NBER Working Paper Series. Cambridge, MA: National Bureau of Economic Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Aronson, E., & Carlsmith, J. M. (1963). Effect of the severity of threat on the devaluation of forbidden behavior. *Journal of Abnormal and Social Psychology*, *66*(6), 584–588.

Atkins-Burnett, S., Fernandez, C., Akers, L., Jacobson, J., & Smither-Wulsin, C. (2012). *Landscape analysis of non-cognitive measures*. Princeton, NJ: Mathematica Policy Research.

Babad, E. Y., Inbar, J., & Rosenthal, R. (1982). Teachers' judgment of students' potential as a function of teachers' susceptibility to biasing information. *Journal of Personality and Social Psychology*, *42*(3), 541–547.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

Berkowitz, M. W. (2012). Moral and character education. In K. R. Harris, S. Graham, T. Urdan, J. M. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook: Vol. 2. Individual differences and cultural and contextual factors* (pp. 247–264). Washington, DC: American Psychological Association.

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, *20*, 821–843.

Binet, A., & Simon, T. (1916). *The development of intelligence in children (The Binet-Simon Scale)*. Baltimore, MD: Williams & Wilkins Co.

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*(3), 899–911.

Blair, C., & Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PLoS ONE*, *9*(11), e112393.

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*(1), 711–731.

Borghans, L., Duckworth, A. L., Heckman, J. J., & ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972–1059.

Bowman, N. A. (2010). Can 1st-year college students accurately report their learning and development? *American Educational Research Journal*, *47*(2), 466–496.

Brackett, M. A., & Geher, G. (2006). Measuring emotional intelligence: Paradigmatic diversity and common ground. In J. Ciarrochi, J. Forgas, & J. D. Mayer (Eds.), *Emotional intelligence in everyday life* (2nd ed., pp. 27–50). New York, NY: Psychology Press.

Brackett, M. A., & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*(9), 1147–1158.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Burgess, P. W. (1997). Theory and methodology in executive function research. In P. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 91–116). East Sussex, UK: Psychology Press.

Campbell, D. T. (1976). *Assessing the impact of planned social change*. Paper presented at the Conference on Social Psychology, Visegrad, Hungary.

Carlson, S. M., Zelazo, P. D., & Faja, S. (2013). Executive function. In P. D. Zelazo (Ed.), *The Oxford handbook of developmental psychology: Vol. 1. Body and mind* (pp. 706–742). New York, NY: Oxford University Press.

Carr, P. B., & Steele, C.M. (2009). Stereotype threat and inflexible perseverence in problem solving. *Journal of Experimental Social Psychology*, *45*, 853–859.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453–484.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122.

Damon, W. (2010). The bridge to character: To help students become ethical, responsible citizens, schools need to cultivate students' natural moral sense. *Education Leadership*, *67*(5), 36–41.

de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*(1), 76–99.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science*, *333*(6045), 959–964.

Diener, E., Larsen, J. E., & Emmons, R. A. (1984). Person × Situation interactions: Choice of situations and congruence response models. *Journal of Personality and Social Psychology*, *47*(3), 580–592.

D'Mello, S., Duckworth, A., & Dieterle, E. (2014). *Advanced, analytic, automated measures of state engagement during learning*. Manuscript under review.

Dobbie, W., & Fryer, R. G., Jr. (2013). *The medium-term impacts of high-achieving charter schools on non-test score outcomes*. NBER Working Paper Series. Cambridge, MA: National Bureau of Economic Research.

Dodge, K. A. (1980). Social cognition and children's aggressive behavior. *Child Development*, *51*, 162–170.

Drucker, P. F. (1974). *Management: Tasks, responsibilities, practices*. New York, NY: Routledge.

Duckworth, A., Gendler, T., & Gross, J. (2014). *Situational strategies for self-control*. Manuscript under review.

Duckworth, A. L. (2009). (Over and) beyond high-stakes testing. *American Psychologist*, *64*(4), 279–280.

Duckworth, A. L., & Carlson, S. M. (2013). Self-regulation and school success. In B. W. Sokol, F. M. E. Grouzet, & U. Muller (Eds.), *Self-regulation and autonomy: Social and developmental dimensions of human conduct* (pp. 208–230). New York, NY: Cambridge University Press.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716–7720.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939–944.

Duckworth, A. L., & Steinberg, L. (2015). Unpacking self-control. *Child Development Perspectives*, *9*(1), 32–37.

Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social Psychological and Personality Science*, *1*(4), 311–317.

Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (Eds.). (2015). *Handbook of social and emotional learning: Research and practice*. New York, NY: Guilford.

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

Easton, J. (2013). *Using measurement as leverage between developmental research and educational practice*. Paper presented at the Center for Advanced Study of Teaching and Learning Meeting, Charlottesville, VA. Retrieved from http://ies.ed.gov/director/pdf/Easton062013.pdf

Egalite, A. J., Mills, J. N., & Greene, J. P. (2014). *The softer side of learning: Measuring students' non-cognitive skills* (EDRE Working Paper No. 2014-03). Fayetteville: University of Arkansas Education Reform.

Eid, M., & Diener, E. (2006). *Handbook of multimethod measurement in psychology* (Vol. 553). Washington, DC: American Psychological Association.

Eisenberg, N., Spinrad, T. L., Fabes, R. A., Reiser, M., Cumberland, A., Shepard, S. A., . . . Thompson, M. (2004). The relations of effortful control and impulsivity to children's resiliency and adjustment. *Child Development*, *75*(1), 25–46.

Elias, M. J. (Ed.). (1997). *Promoting social and emotional learning: Guidelines for educators*. Chicago, IL: Association for Supervision and Curriculum Development.

Ent, M. R., Baumeister, R. F., & Tice, D. M. (2015). Trait self-control and the avoidance of temptation. *Personality and Individual Differences*, *74*, 12–15.

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

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28.

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 98–143). Hoboken, NJ: Jossey-Bass.

Fleeson, W., & Noftle, E. E. (2008). The end of the person-situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass*, *2*(4), 1667–1684.

Freud, S. (1920). *Introductory lectures on psychoanalysis*. New York, NY: Norton.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review*, *15*(4), 352–366.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182.

Galla, B. M., & Duckworth, A. L. (2015). More than resisting temptation: Beneficial habits mediate the relationship between self-control and positive life outcomes. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspp0000026

Galla, B. M., Plummer, B. D., White, R., Meketon, D., D'Mello, S. K., & Duckworth, A. L. (2014). Development and validation of the Academic Diligence Task. *Contemporary Educational Psychology*, *39*(4), 314–325.

Gardner, H. (2004). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*(4), 380–387.

Goldhaber, D., & Loeb, S. (2013). *What do we know about the tradeoffs associated with teacher misclassificationin high stakes personnel decisions?* Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/teacher-misclassifications/

Goldman, S. (2006). Self-discipline predicts academic performance among low-achieving adolescents. *Res: A Journal of Undergraduate Research*, *2*(1), 84–97.

Greenberg, M. T. (2010). School-based prevention: Current status and future challenges. *Effective Education*, *2*(1), 27–52.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*(4), 495–525.

Hartshorne, H., & May, M. A. (1929). *Studies in the nature of character: Studies in self-control* (Vol. 2). New York, NY: McMillan.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385.

Heckman, J. J., & Kautz, T.D. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464.

Heckman, J. J., & Kautz, T. D. (2014a). Achievement tests and the role of character in American life. In J. J. Heckman, J. E. Humphries, & T. Kautz (Eds.), *The myth of achievement tests: The GED and the role of character in American life* (pp. 1–71). Chicago, IL: University of Chicago Press.

Heckman, J. J., & Kautz, T. D. (2014b). *The myth of achievement tests: The GED and the role of character in american*. Chicago, IL: University of Chicago Press.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918.

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B., Spinell, A., Guare, J., & Rohrbeck, C. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review*, *15*(3), 393–409.

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, *102*(6), 1318.

Hollingworth, L., Dude, D. J., & Shepherd, J. K. (2010). Pizza parties, pep rallies, and practice tests: Strategies used by high school principals to raise percent proficient. *Leadership and Policy in Schools*, *9*(4), 462–478.

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., . . . Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish–little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170.

Imhoff, R., Schmidt, A. F., & Gerstenberg, F. (2013). Exploring the interplay of trait self-control and ego depletion: Empirical evidence for ironic effects. *European Journal of Personality*, *28*(5), 413–424.

Inzlicht, M., & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of Personality and Social Psychology*, *99*(3), 467–481.

Inzlicht, M., McKay, L., & Aronson, J. (2006). Stigma as ego depletion: How being the target of prejudice affects self-control. *Psychological Science*, *17*(3), 262–269.

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, *99*(3), 549–571.

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina*. NBER Working Paper Series. Cambridge, MA: National Bureau of Economic Research.

Jackson, J. J., Connolly, J. J., Garrison, M., Levine, M., & Connolly, S. L. (2015). Your friends know how long you will live: A 75 year study of peer-rated personality traits. *Psychological Science*, *26*(3), 335–340.

Jones, E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349–364.

Kane, T. J., & Cantrell, S. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Seattle, WA: Bill and Melinda Gates Foundation.

Kautz, T. D., & Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. Unpublished manuscript, Department of Economics, University of Chicago, Chicago, IL.

Kelvin, W. T. (1883). *Popular lectures and addresses* (Vol. 1). London, UK: MacMillan.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., . . . Seligman, M. E. P. (2014). From "Sooo excited!!!" to "So proud": Using language to study development. *Developmental Psychology*, *50*(1), 178–188.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*(1), 109–114.

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

King, M. L., Jr. (1947, January/February). The purpose of education. *Maroon Tiger*.

Kovacs, M. K. (1992). *Children's Depression Inventory–Short Form (CDI)*. New York, NY: Multi-Health Systems.

Kristjánsson, K. (2013). Ten myths about character, virtue and virtue education—plus three well-founded misgivings. *British Journal of Educational Studies*, *61*(3), 269–287.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

Krosnick, J. A., & Fabrigar, L. R. (in press). *The handbook of questionnaire design*. New York, NY: Oxford University Press.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263–314). Bingley, UK: Emerald Group.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis*, *52*(2), 161–193.

Kyllonen, P. C., & Bertling, J. (Eds.). (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis: Background, technical issues, and methods of data analysis* (pp. 277–286). London, UK: Chapman Hall/CRC Press.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. San Francisco, CA: Jossey-Bass.

Lapsley, D. K., & Yeager, D. S. (2012). Moral-character education. In I. B. Weiner, W. M. Reynolds, & G. E. Miller (Eds.), *Handbook of psychology: Vol. 7. Educational psychology* (2nd ed., pp. 289–348). New York, NY: Wiley.

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, *48*(6), 1291–1299.

Levin, H. M. (2013). The utility and need for incorporating noncognitive skills into large-scale educational assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 67–86). New York, NY: Springer Netherlands.

Lucas, R. E., & Baird, B. M. (2006). Global self-assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29–42). Washington, DC: American Psychological Association.

Lucas, R. E., & Lawless, N. M. (2013). Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments. *Journal of Personality and Social Psychology*, *104*(5), 872–884.

Marat, D. (2005). Assessing mathematics self-efficacy of diverse students from secondary schools in Auckland: Implications for academic achievement. *Issues in Educational Research*, *15*(1), 37–68.

McGuire, J. T., & Kable, J. W. (2013). Rational temporal predictions can underlie apparent failures to delay gratification. *Psychological Review*, *120*(2), 395–410.

Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science*, *21*(4), 539–541.

Merrell, K. W., & Gueldner, B. A. (2010). *Social and emotional learning in the classroom: Promoting mental health and academic success*. New York, NY: Guilford Press.

Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, *71*(3), 281.

Mischel, W. (1961). Father-absence and delay of gratification. *Journal of Abnormal and Social Psychology*, *63*(1), 116–124.

Mischel, W. (1968). *Personality and assessment*. Hoboken, NJ: Wiley.

Mischel, W. (2014). *The Marshmallow Test: Mastering self-control*. New York, NY: Little, Brown.

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., . . . Zayas, V. (2011). 'Willpower' over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*(2), 252–256.

Mischel, W., & Liebert, R. M. (1967). The role of power in the adoption of self-reward patterns. *Child Development*, *38*(3), 673–683.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H. L., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693–2698.

Müller, U., Kerns, K. A., & Konkin, K. (2012). Test–retest reliability and practice effects of executive function tasks in preschool children. *Clinical Neuropsychologist*, *26*(2), 271–287.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. New York, NY: Free Press.

Murphy, M. C., & Dweck, C. S. (2009). A culture of genius: How an organization's lay theory shapes people's cognition, affect, and behavior. *Personality and Social Psychology Bulletin*, *36*(3), 282–296.

Naemi, B., Burrus, J., Kyllonen, P. C., & Roberts, R. D. (2012, December). *Building a case to develop noncognitive assessment products and services targeting workforce readiness at ETS*. Princeton, NJ: Educational Testing Service.

National Institute of Child Health and Human Development. (1999). *Child's self-regulation fifty-four month delay of gratification test*. Research Triangle Park, NC: Author.

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count*. New York, NY: Norton.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*(2), 130–159.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

O'Brien, J., Yeager, D. S., Galla, B., D'Mello, S., & Duckworth, A. L. (2015). *Between-school comparisons in non-cognitive factors: Evidence of reference bias in self-reports and advantages of performance tasks*. Manuscript in preparation.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187–207.

Pace, C. R., & Friedlander, J. (1982). The meaning of response categories: How often is "occasionally," "often," and "very often"? *Research in Higher Education*, *17*(3), 267–281.

Park, A., Tsukayama, E., Patrick, S., & Duckworth, A. L. (2015). *A tripartite taxonomy of character*. Manuscript in preparation.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academy of Sciences.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. Washington, DC: American Psychological Association.

Pickering, A. D., & Gray, J. A. (1999). The neuroscience of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 277–299). New York, NY: Guilford Press.

Ployhart, R. E., & MacKenzie, W. I., Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Vol. 2. Selecting and developing members for the organization* (pp. 237–252). Washington, DC: American Psychological Association.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76*(1), 85–97.

Raudenbush, S. W. (2013). *What do we know about using value-added to compare teachers who work in different schools?* Stanford, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/08/CKN_Raudenbush-Comparing-Teachers_FINAL_08-19-13.pdf

Raudenbush, S. W., & Jean, M. (2012). *How should educators interpret value-added scores?* Stanford, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/

Raver, C. C., Jones, S. M., Li-Grining, C. P., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, *82*(1), 362–378.

Ravitch, D. (2012). *What is Campbell's law?* Retrieved from http://dianeravitch.net/2012/05/25/what-is-campbells-law/

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*(5), 535–549.

Revelle, W., Wilt, J., & Rosenthal, A. (2010). *Individual differences in cognition: New methods for examining the personality-cognition link*. New York, NY: Springer Science.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*(1), 3–25.

Roberts, B. W., Jackson, J. J., Duckworth, A. L., & Von Culin, K. (2011). Personality measurement and assessment in large panel surveys. *Forum for Health Economics and Policy*, *14*(3), 1–32.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1–25.

Roberts, R. D., Markham, P. M., Matthews, G., & Zeidner, M. (2005). Assessing intelligence: Past, present, and future. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 333–360). Thousand Oaks, CA: Sage.

Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 3–50). Hoboken, NJ: Wiley.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York, NY: McGraw-Hill.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*(1), 18–38.

Sabini, J., Siepmann, M., & Stein, J. (2001). The really fundamental attribution error in social psychological research. *Psychological Inquiry*, *12*(1), 1–15.

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance*, *20*(3), 179–185.

Sackett, P. R. (2011). Faking in personality assessment: Where do we stand? In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 330–344). Oxford, UK: Oxford University Press.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 61–79.

Schmidt, F. L. (April, 2013). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research*. Paper presented at the University of Iowa, Iowa City.

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

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York, NY: Academic Press.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, *22*(2), 127–160.

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*(2), 374–408.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013) *Measuring 21st century competencies: Guidance for educators*. Santa Monica, CA: RAND Corporation.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, *94*(4), 718.

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan*. Santa Monica, CA: RAND Corporation.

Sternberg, R. J. (2008). Using cognitive theory to reconceptualize college admissions testing. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (pp. 159–175). New York, NY: Lawrence Erlbaum.

Tough, P. (2011, September 14). What if the secret to success is failure? *New York Times Magazine*, pp. 1–14.

Tough, P. (2013). *How children succeed: Grit, curiosity, and the hidden power of character*. New York, NY: Houghton Mifflin Harcourt.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Tsukayama, E., Duckworth, A. L., & Kim, B. E. (2013). Domain-specific impulsivity in school-age children. *Developmental Science*, *16*(6), 879–893.

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes*. Washington, DC: Mathematica Policy Research.

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243–262.

Wagerman, S. A., & Funder, D. C. (2007). Acquaintance reports of personality and academic achievement: A case for conscientiousness. *Journal of Research in Personality*, *41*(1), 221–229.

Wagerman, S. A., & Funder, D. C. (2009). Personality psychology of situations. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 27–42). Cambridge, UK: Cambridge University Press.

Weissberg, R. P., & Cascarino, J. (2013). Academic learning + social-emotional learning = national priority. *Phi Delta Kappan*, *95*(2), 8–13.

Weschler, D. (1943). Non-intellective factors in general intelligence. *Journal of Abnormal and Social Psychology*, *38*(1), 101–103.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2015). *Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling*. Manuscript under review.

White, J. L., Moffitt, T. E., Caspi, A., Bartusch, D. J., Needles, D. J., & Stouthamer-Loeber, M. (1994). Measuring impulsivity and examining its relationship to delinquency. *Journal of Abnormal Psychology*, *103*(2), 192–205.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. New York, NY: College Entrance Examination Board.

Wong, M. M., & Csikszentmihalyi, M. (1991). Affiliation motivation and daily experience: Some issues on gender differences. *Journal of Personality and Social Psychology*, *60*(1), 154–164.

Yeager, D. S., & Bryk, A. S. (2014). *Practical measurement*. Unpublished manuscript, Department of Psychology, University of Texas at Austin, Austin.

Yeager, D. S., Henderson, M., Paunesku, D., Walton, G. M., D'Mello, S., Spitzer, B. J., & Duckworth, A. L. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. *Journal of Personality and Social Psychology*, *107*(4), 559–580.

Yeager, D. S., & Krosnick, J. (2011). Does mentioning "some people" and "other people" in a survey question increase the accuracy of adolescents' self-reports? *Developmental Psychology*, *47*(6),1674–1679.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301.

Ziegler, M., MacCann, C., & Roberts, R. (Eds.). (2011). *New perspectives on faking in personality assessment*. Oxford, UK: Oxford University Press.

Zirkel, S., Garcia, J. A., & Murphy, M. C. (2015). Experience-sampling research methods and their potential for education research. *Educational Researcher, 44*, 7–16. doi: 10.3102/0013189x14566879

AUTHORS

ANGELA L. DUCKWORTH, PhD, is an associate professor of psychology at the University of Pennsylvania, 3701 Market St., Suite 215, Philadelphia, PA 19104; *duckwort@psych.upenn.edu*. Her research focuses on non-IQ competencies, including self-control and grit, which predict success both academically and professionally.

DAVID SCOTT YEAGER, PhD, is an assistant professor of psychology at the University of Texas at Austin, 108 E. Dean Keeton Stop A8000, Austin, TX, 78712-1043, and a fellow at the Carnegie Foundation for the Advancement of Teaching; *dyeager@utexas.edu*. His research focuses on adolescent development, psychological intervention, and psychological measurement.

Manuscript received June 10, 2014
Revisions received January 7, 2015, and March 19, 2015
Accepted March 23, 2015