19

Assessment of Social-Emotional, Soft, Character, Behavioral, and Intrapersonal and Interpersonal Skills

Patrick C. Kyllonen
ETS

Jiyun Zu
ETS

Kyllonen, P. C., & Zu, J. (2025). Assessment of social-emotional, character, behavioral, and intrapersonal and interpersonal skills. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1339–1421). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0019

> Intelligence plus character—that is the goal of true education.
> Martin Luther King Jr.

We have long recognized that schools develop character along with mathematics and language skills. Students are taught, at least implicitly, to pay attention to the instructor, to put effort into their work, to show up and complete assignments on time, to be considerate of other students, to avoid arguing with the teacher and fighting with other students, to be honest and truthful, and to develop a love of learning. As a society, we believe that such behaviors and attitudes are important for the future success of students—success in education, in the workforce, and in life, generally. However, these beliefs have not typically translated into policy or practice in educational testing.

Testing reflects policy priorities and influences what is taught and learned (Frederiksen, 1984; Levin, 2012). Since 2001, federal and state policies have centered around evaluation of students' knowledge of subject areas, as seen in the Every Student Succeeds Act and its predecessor, the No Child Left Behind Act of 2001, as well as in our *Nation's Report Card* (National Assessment of Educational Progress). International assessments such as the Program for International Assessment (PISA), which began in 2000; Trends in International Math and Science Study (TIMSS), beginning in 1995; and Progress in International Reading Literacy Study (PIRLS), beginning in 2001, along with national accountability assessments in most countries of the Organisation for Economic Co-operation and Development (OECD) all focus on curricular content knowledge. We evaluate teachers with value-added models based entirely on student gains in academic achievement, as measured by curricular tests (McCaffrey et al., 2003).

Only recently has there been an explicit attempt to move beyond a general acknowledgment of the school's role in building character to embracing character development as part of the core school mission warranting its monitoring. In the United States, a subset of 10 California school districts (the California Office to Reform Education) received a waiver from the federal government to implement their own accountability assessment, which included social-emotional learning skills (West, 2016). The District of Columbia Public Schools is administering social and emotional learning (SEL) survey measures to assess readiness (District of Columbia Public Schools, 2017). Many states have by now published social-emotional learning standards paralleling mathematics, English language arts, and science standards (Collaborative for Academic, Social, and Emotional Learning [CASEL], 2020a). Chile's ministry of education is mandated to include social-emotional skills as part of its national accountability assessment (Ministerio de Educacio, 2016); other countries are following suit (Bravo-Senzana et al., 2023). The OECD recently launched its Study on Social and Emotional Skills parallel to PISA, but designed to monitor progress in student development of social-emotional skills, rather than cognitive skills (OECD, 2019a). Many organizations have issued position papers on the importance of SEL (Aspen Institute, 2018; Atwell & Bridgeland, 2019; OECD, 2015; Salzburg Global Seminar, 2018; Schanzenbach et al., 2016). Online resources are now available to find and evaluate

social-emotional learning assessments, including Rand's (Hamilton et al., 2018) Education Assessment Finder, CASEL's (2020b) SEL Assessment Guide, California's Guide to Social and Emotional Learning Resources (California Department of Education, 2018), and others. What has changed?

EVIDENCE FOR THE IMPORTANCE OF INTERPERSONAL AND INTRAPERSONAL SKILLS

Benefits-of-Education Studies

One important source for change was economists working from a human capital theory framework. Human capital theory (Becker, 1994) is the idea that individuals are rewarded for their productivity (e.g., measured as earnings, or labor market outcomes more generally, including employment and nonincarceration), which is driven by the set of skills, knowledge, and health individuals bring to the labor market. These attributes are affected by education, on-the-job training, medical care, and other factors. This framework enables quantifying the value of investments in education on economic productivity, the rate of return for a high school or college education (Card, 1999). A finding is that educational attainment is an important causal determinant of employment and earnings (private returns to education; Barrow & Rouse, 2005; Card, 1999) as well as national economic strength (social returns to education; Moretti, 2005), hence the public investment in education. However, the benefit of education on labor market outcomes is only partly due to the gains in cognitive skills associated with education, perhaps only about 20% (Bowles et al., 2001). Most of the benefit is due to factors other than the cognitive skills acquired, as measured with standardized tests, hence the attribution to noncognitive skills. Recent evidence shows that returns to cognitive ability have even gone down since the 1980s, with a 30% to 50% larger effect of wages in the 1980s compared to the 2000s (Castex & Dechter, 2014).

Another key demonstration of the benefits of education not accounted for by cognitive ability was Heckman and Rubenstein's (2001) finding that General Educational Development (GED) holders, individuals who failed to complete high school but scored comparably to high school degree holders on cognitive tests and therefore were similar in cognitive skills, experienced very different labor market outcomes. GED holders were far more likely to be unemployed and have trouble with the law (Heckman & Rubinstein, 2001; Heckman et al., 2014). The signaling value of a high school diploma is a potential alternative explanation for these findings—employers value degrees—but empirical studies support a human capital interpretation over a signaling interpretation (Clark & Martorell, 2014).

A third source of findings concerns the value of early childhood education. Although such programs were motivated with the goal of boosting cognitive ability for children living in impoverished environments, a common finding was that cognitive ability gains tended to fade out after a few years, as determined by comparison with control groups (Bailey et al., 2017). However, long-term treatment effects have consistently

been realized in many positive outcomes, including increased educational attainment and employment and less incarceration (Barnett, 1996).

The significance of all three of these study clusters is that they demonstrate the importance of noncognitive skills—skills associated with education that were not related to the cognitive gains education produces. However, aside from the cognitive versus noncognitive distinction, these studies tended not to identify the specific nature of the noncognitive factors associated with the education effect. A question is, What are these "noncognitive" benefits of education?—What is it that teachers teach and students learn, other than mathematics, reading, science, and other subject matters, that are of value in school, to employers, and in life?

Predictions-From-Measures Studies

Studies have been conducted by economists and psychologists examining the predictions of educational, workplace, and life outcomes from self- or other-report surveys. In higher education, a meta-analysis of psychosocial and study skill factors by Robbins et al. (2004) found the highest correlates of grades to be academic self-efficacy and achievement motivation and those of retention to be academic goals, academic self-efficacy, and academic-related skills. These factors were predictive after controlling for socioeconomic status, cognitive ability (test scores), and high school grades. Richardson et al. (2012) similarly found academic self-efficacy to be among the highest correlates of grades, along with effort regulation, need for cognition, and grade goal. In economics, Lindqvist and Vestman (2011) found that a composite measure based on ratings by a clinical psychologist of 18-year-old males predicted low earnings and chronic unemployment 20 years later. Segal (2013) found that misbehavior ratings of eighth-grade students by their teachers predicted earnings 20 years later after controlling for educational attainment and cognitive test scores.

These studies are suggestive of the importance of noncognitive factors but were also limited by the lack of a reliable strategy for categorizing noncognitive predictors. Beginning in the late 1980s and into the 1990s, the Big Five or five-factor model of personality arose in prominence (P. T. Costa & McCrae, 1992; Digman, 1990; Goldberg, 1993; John & Srivastava, 1999). We discuss this model further in the section "What Are the Key Skills?" but for our purposes here, it suffices to say that the model posits that individual personality descriptions are captured by five independent dimensions, Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. As in many realms, particularly measurement, standardization leads to progress (Cochrane, 1966), and the adoption of the five-factor model had a major impact on the value and frequency of prediction studies. Meta-analyses of workplace outcomes (Barrick & Mount, 1991; Dudley et al., 2006; Rothmann & Coetzer, 2003; Salgado, 1997; Tett et al., 1991), lifetime earnings outcomes (Gensowski, 2018), general life outcomes (B. W. Roberts et al., 2007), and educational outcomes (Noftle & Robins, 2007; Poropat, 2009; Salgado & Tauriz, 2014) have been conducted within the Big Five framework, leading to findings particularly of the importance of Conscientiousness (i.e.,

being organized, responsible, and achievement striving) on school, life, and workforce outcomes (Almlund et al., 2011; Levin, 2012; National Research Council, 2012; Ozer & Benet-Martinez, 2006).

Note on the Relative Importance of Cognitive Versus Noncognitive Skills

Cognitive ability is sometimes treated as a control variable to examine the incremental predictiveness of noncognitive variables (Noftle & Robins, 2007; Robbins et al., 2004; Segal, 2013). However, several studies have also compared the prediction given by cognitive versus noncognitive variables. Poropat (2009) and Richardson et al. (2012) found that some of the noncognitive factors have as strong a relationship with outcomes as cognitive abilities do. Poropat (2009) computed correlations of academic performance with intelligence, $\rho = .25$ (47 studies, all levels), and Conscientiousness, $\rho = .22$ (138 studies, all levels), findings in line with findings from the workforce (Levin, 1989; Salgado & Tauriz, 2014). Borghans et al. (2016) examined four large longitudinal data sets to show that personality, as expressed in grades and achievement scores, was more predictive of various life outcomes (e.g., wages, education, arrests, life satisfaction, civic participation) than IQ scores were. B. W. Roberts et al. (2007) found that the correlations between personality measures and a variety of life outcomes (mortality, divorce, occupational attainment) were of the same magnitude as the correlations between cognitive ability measures and socioeconomic status measures with those outcomes.

One complication in reporting simple correlations is correcting for range restriction—Levin (2012) pointed out that such corrections can distort findings. Selection on cognitive and noncognitive ability can operate differently—selecting on cognitive ability is often direct and affects the variance of cognitive test scores on the selected sample. Selecting on noncognitive factors is often indirect and often unmeasured. However, a persistent finding is that within a cohort, the correlation between noncognitive scores and outcomes is stronger for those in the lower half of the ability distribution (Lindqvist & Vestman, 2011; Segal, 2013). Noncognitive ability appears to be particularly important for students who leave school early and have difficulties maintaining employment.

A challenge in comparing the relative importance of cognitive and noncognitive skills is that measures of them call on both skills so their effects are difficult to disentangle (Borghans et al., 2011, 2016; Kyllonen & Kell, 2018). Being unmotivated temporarily or characteristically can produce low cognitive test scores and comprehension difficulties can affect responses on a personality survey.

Workplace-Demands Studies

Employer surveys ask employers what skills they look for in new graduates or how important various skills were for job performance. The findings from such surveys have been consistent. Casner-Lotto and Barrington (2006) found that what they called *applied skills* (e.g., critical thinking/problem-solving, oral and written communications, teamwork/collaboration, diversity, leadership, professionalism/work ethic) were

far more likely to be considered "very important" to success at work by employers compared to *basic knowledge/skills* (e.g., English language, reading comprehension, mathematics, science), particularly for high school and 2-year graduates. For these two groups, the two skills most likely to be rated as most important were professionalism/work ethic and teamwork/collaboration. Other surveys have found similar results, both domestically (Cengage, 2019; National Association of Colleges and Employers, 2018) and internationally (Barton et al., 2012).

A question is whether employer surveys provide a distorted picture of the skills sought. Employers might say *X* is important when asked in a survey, while looking for *Y* during recruiting. Rios et al. (2020) examined 142,000 job advertisements and found that 70% of job ads requested at least one of what they called "21st-century skills," with the top skills being oral and written communication, collaboration, problem-solving, communication skills, social intelligence, and self-direction. These results are largely in line with results from the employer surveys, although terminology differences make them somewhat hard to compare. Employers routinely assess these skills using personality tests, structured interviews, situational judgment tests, game-based assessments, and reference checking (U.S. Office of Personnel Management, 2018b).

Another kind of workplace demand study has involved examining the actual activities workers engage in on the job. These studies typically have the theme of effects of technology on the nature of work. An influential early study by Autor et al. (2003) examined jobs (using the Dictionary of Occupational Titles database) in the years 1960 to 1998 and found that the effect of technology was to substitute for routine cognitive and manual tasks but complement activities involving nonroutine analytical and interpersonal tasks. This has led to a decline in manual tasks (routine and nonroutine) and a rise in nonroutine analytical and interpersonal tasks, thereby increasing the demand for the latter. The analysis was later extended to 2009, showing even clearer trends, but also showing a plateauing of nonroutine tasks (Autor & Price, 2013). Deming (2017) conducted a similar study covering the years 1980 to 2012 and showed that jobs requiring social interactions have grown considerably, showing a trend for an increasing "social skills" premium. Weinberger (2014) presented similar findings using Dictionary of Occupational Titles data showing a growth in employment and earnings for occupations requiring high levels of both cognitive and social skills compared to those requiring only one or the other. Burrus et al. (2013) applied a principal component analysis to O*NET ratings, identifying 15 competency components (mixtures of cognitive and noncognitive competencies), and showed that the most important competencies by both importance ratings and correlations with earnings were what they labeled problem-solving, fluid intelligence, teamwork, achievement/innovation, and communication skills.

There also have been studies attempting to project skill demands for the workplace of the future. A workshop conducted by the National Research Council (2008) reinforced the findings of the growing importance of broad interpersonal and social skills, noting the same conclusion in the earlier SCANS Commission report (Kane

et al., 1990) and suggesting the addition of adaptability. A study by Frey and Osborne (2017) attempted to predict the probability of computerization for 702 occupations in the United States, based on O*NET data, concluding that close to half of U.S. employment was at risk. They showed that the most resistant occupations were ones involving perception (finger and manual dexterity), creativity (originality and fine arts), and social intelligence (social perceptiveness, negotiation, persuasion, and caring for others), particularly when in high-wage jobs requiring high levels of educational attainment. A study by McKinsey Global Institute (Manyika et al. (2017) concluded that future occupations will require more interactions and management, as well as more social-emotional, creative, and logical reasoning abilities. A recent edited volume by Oswald et al. (2019) explored educational implications for the future of work (Hilton, 2019).

These three clusters of studies—benefits of education, predictions from measures, and workplace demands—all point to a growing recognition of the importance of noncognitive, character, or social-emotional skills. There also appears to be evidence that the importance of these skills is likely to increase in the future as a result of changes in the workforce related to technology. This highlights the importance of the topic for educational measurement.

A Note on Terminology

In the title of this chapter, we refer to interpersonal and intrapersonal skills and other names; in the chapter, we use various terms to refer to the same general set of attributes. Terminology in this field is a source of frustration and confusion (Duckworth & Yeager, 2015; Reeves & Venator, 2014). SEL or social-emotional competencies or skills is terminology that has taken hold in K–12 education circles (CASEL, 2020b). It perhaps supersedes character or character skills, although character education is still a prominent reference. For a long time, What Works Clearinghouse maintained a separate Character Education category (Institute of Educational Sciences, 2007), although that appears to now be replaced by a Behavior category (Institute of Educational Sciences, 2020). In higher education, interpersonal and intrapersonal skills are widely used because of the National Academies of Science's (2017) adoption of that terminology. Also used is "hard-to-measure" skills or competencies (Stecher & Hamilton, 2014) and 21st-century competencies (Soland et al., 2013). In the workforce, soft skills (Kyllonen, 2013) and behavioral competencies (Society for Human Resource Management, 2014) are widely used, as is emotional intelligence (Goleman, 1995) (Emotional intelligence also has a more technical meaning; J. D. Meyer et al., 2008.) Economists have historically referred to noncognitive skills, to contrast with the cognitive skills measured with IQ and standardized achievement tests (although Messick, 1978, used the term noncognitive, noting its problems, but blaming its necessity on Bloom's 1956 usurpation of cognitive to define a particular set of skills). Industrial-organizational psychology refers to "other factors" as seen in the acronym KSAO, for knowledge, skills, abilities, and other factors.

Skills are prominent in economic or human capital theory discussions as a shorthand for whatever is learned in school or in training that has value in the workforce. But measurement psychologists have historically distinguished knowledge, skills, and abilities and added attitudes, values, beliefs, and behaviors, as seen in Buros Mental Measurements Yearbook (Carlson et al., 2020). Twenty-first-century skills, skills for the new economy, and others appear in the literature to reflect the changing demands of school and the workplace as a result of changes in technology (Oswald et al., 2019). *Skill* itself as a term may be thought of as controversial to describe some attributes, but its evolving widespread acceptance is shown in a *New York Times* opinion piece by U.S. senator Jeff Flake (2017), titled "We Need Immigrants With Skills, But Working Hard Is a Skill." Personality is also used. However, the concept of personality, like the concept of intelligence, is associated with heredity and permanence, and so educators have tended to shy away from this terminology, believing personality to be outside the control of teachers and the educational system. As discussed in the next section, the foundation for this belief may be chipping away.

MALLEABILITY OF PERSONALITY

A recurring problem in the field of the assessment of noncognitive skills in education is the widespread belief that personality or character traits are fixed, and therefore outside the realm of education. Personality traits have been defined as the "relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances" (B. W. Roberts, 2009). Numerous studies have shown that personality traits are rank-order stable over decades (Ferguson, 2010) and heritable (Vukasovic & Bratko, 2015). However, this is true of achievement as well (Rimfeld et al., 2018), and also of social cognitive constructs, such as interests, self-concept, and academic effort, popularly believed to be more malleable, but which have been found to be no more so than the Big Five personality traits (Rieger et al., 2017).

Damian et al. (2019) tracked personality change in 1,795 adolescents tested 50 years later, finding moderate stability of personality across that period ($r = .31$ corrected for measurement error). But they also found growth of about half a standard deviation for personality factors over the period, in the direction of maturity (over time, greater Emotional Stability, Conscientiousness, Agreeableness, Openness).[1] Specific activities, such as psychotherapy, can have relatively swift and substantial effects on personality change (B. W. Roberts et al. 2017). Social-emotional learning interventions in school have similar and lasting effects on students' personality as well as academic and other outcomes (Durlak et al., 2010, 2011; Taylor et al., 2017). Together, these findings challenge beliefs about the permanence of noncognitive factors or their inaccessibility to education, whether they are called personality traits, social-emotional competencies, noncognitive skills, or something else. Like habits, personality can change; personality is relatively enduring but not fixed.

WHAT ARE THE KEY SKILLS?

Considerable evidence supports the importance of noncognitive skills in school, the workplace, and life. What are those skills? What noncognitive skills are most important, and what skills are the ones driving educational and workforce outcomes? What skills should be given attention in education, and what skills should be monitored for growth and development?

Two specific skills have received an enormous amount of attention in education circles. One is grit (Duckworth, 2016), and the other is growth mindset (Dweck, 2006). Since 2016, it seems to have been almost impossible to attend a meeting of educational practitioners or policy makers without the word *grit* being mentioned. Grit seems to play an outsized role, serving as a stand-in for all the skills that are missed by standardized cognitive tests. Growth mindset has played a similar role and has received a comparable number of mentions in practitioner and policy discussions. A solid scientific literature attests to the importance of the constructs of grit and growth mindset on outcomes. However, something else may be at play. Both constructs have simple but compelling narratives—for grit, work hard in a sustained way and you will succeed; for growth mindset, believe that your successes and failures are a result of the effort put in, rather than your intelligence or external factors, and by doing so you will respond more productively to successes and failures ("the power of yet"). The two constructs also have charismatic advocates with popular TED talks (Duckworth, 2013, 50 languages, over 19 million views; Dweck, 2014, 42 languages, over 10 million views) and best-selling books (Duckworth, 2016, 21 weeks on the *New York Times* bestseller list; Dweck, 2017).

There have been meta-analyses of the constructs. Credé et al. (2017) examined 584 effect sizes from 88 samples and 66,807 individuals and showed that Duckworth's grit was uncorrelated with cognitive ability ($\rho = .05$)[2] and highly correlated with Conscientiousness from the Big Five ($\rho = .84$). Grit predicted grades, both in high school ($\rho = .16$) and in college ($\rho = .17$; the study did not report grit's incremental prediction beyond general cognitive ability, but given the low correlation between grit and cognitive ability, the result likely would have been similar). Grit could be separated into two correlated constructs, perseverance and consistency (correlated $\rho = .66$), and perseverance was shown to be the stronger correlate of grades ($\rho = .26$ and $\rho = .10$, respectively). These findings are largely in line with Poropat's (2009) findings and others reviewed in the previous section.

Several meta-analyses of the growth mindset work have been conducted. Sisk et al. (2018) examined $k = 273$ studies with over $N = 365{,}000$ students and showed that the correlation between growth mindset and achievement was $\rho = .12$, and only about 37% of the studies found a correlation greater than zero. They also found that this relationship was not moderated by academic risk, socioeconomic status, or test score versus grades outcomes. But it was moderated by age, such that the correlation was zero for adults. Sisk et al. (2018) also examined the effect size of mindset interventions, finding $k = 43$ studies ($N = 57{,}155$ participants) attempting to improve growth mindset. Of the 43 studies, 5 of them showed significant improvement, with an estimated effect size of

d = .08. Age, at-risk status, intervention type, and intervention length did not moderate effect size. But there were indications that lower socioeconomic groups benefited more and that reading-based interventions were more effective than computer programs. Yeager et al.'s (2019) randomized control trial (N = 6,320 relatively low-achieving students) found a similar estimate, d = .11, on grades. Lazowski and Hulleman (2016; k = 92, N = 38,377) estimated a much higher d = .49, although they included both randomized and quasi-experimental designs and a larger group of "motivation interventions." Other than the latter estimate, these effect sizes are "small" by Cohen's (1992) rules of thumb, although Yeager et al. (2019) argued that d = .2 should be considered a large effect and that their intervention attained a substantial proportion of that effect.

However, as can be seen from Hattie's (2009) review of effect sizes related to student achievement, there may be many more potentially important factors than grit and growth mindset. Of Hattie's 252 effect sizes, 225 of these showed $|d| > .11$ (Yeager et al.'s 2019 estimate is based on a randomized control trial, whereas Hattie's also included observational data).

At the time of this writing, the CASEL SEL Assessment Guide (2020b) listed 157 constructs from 26 assessments. An organizing framework is clearly needed. John and De Fruyt (2015) provided a way to organize all "21st-century constructs" identified in a review (Trilling & Fadel, 2009). Table 19.1 provides the results of a principal component analysis of self-ratings on these constructs by 350 University of California undergraduates. Five components were extracted that align with the Big Five, suggesting that the five-factor model is a robust characterization of human self-descriptions, regardless of the origins of those self-descriptions.

MOST IMPORTANT SKILLS TO ASSESS: FRAMEWORKS

Various proposals suggest how best to organize the large pool of noncognitive constructs and to highlight the most important ones. We review those that have the most impact on education discussions.

Collaborative for Academic, Social, and Emotional Learning (CASEL): Five-Dimension Framework

CASEL has influenced policy and practice related to social and emotional skills in schools in the United States, particularly in pre-K–12. The CASEL website states that "CASEL was formed in 1994 with the goal of establishing high-quality, evidence-based social and emotional learning (SEL) as an essential part of preschool through high school education." CASEL proposes (a) to conduct research on the efficacy of SEL, (b) to assist in practice and implementation, and (c) to promote legislation at the local, state, and federal levels.

CASEL worked with the Illinois State Board of Education to implement SEL student learning standards for three broad goals of (a) developing self-awareness and

Table 19.1 Principal Component Analysis of 21st-Century Skills and Big Five Alignment

Factor 1: Collaboration (A)	Factor 2: Task Performance (C)	Factor 3: Emotion Regulation (N)	Factor 4: Engagement With Others (E)	Factor 5: Open-mindedness (O)
Compassion	Self-discipline	Self-confidence	Social connection	Curiosity
Care	Focus	Self-esteem	Teamwork	Inquisitiveness
Cooperation	Perseverance	Decisiveness	Social awareness	Willingness to try new ideas
Kindness	Self-control, school	Tackling tough problems	Public speaking	Receptivity
Respect for others	Grit	Cheerfulness	Assertiveness	Innovation
Empathy	Organization	Happiness	Leadership	Vision
Tolerance	Diligence	Optimism	Courage	Insight
Fairness	Precision	Tranquility	Charisma	Tinkering (inventing)
Trust	Dependability	Balance	Speaking out/taking a stand	Learning from mistakes
Forgiveness	Reliability	Stability	Bravery	Excitement creating something new
Gratitude	Consistency	Equanimity	Enthusiasm	Appreciating beauty in the world
Appreciation of others	Trustworthiness	Self-compassion	Passion	Living in harmony with nature
Living in harmony with others	Goal orientation	Self-kindness	Zeal	Spirituality
Interconnectedness	Motivation		Inspiration	Mindfulness
Inclusiveness	Work ethic		Spunk	Existentiality
	Effort		Spontaneity	Awe
	Productivity		Playfulness	Wonder
			Humor	Reverence
				Self-reflection
				Self-awareness
				Consciousness
				Self-actualization
				Authenticity

Note. Column headings indicate aligning Big Five factor: A = Agreeableness, C = Conscientiousness, N = Neuroticism, E = Extraversion, O = Openness. Adapted from "*Education and Social Progress: Framework for the Longitudinal Study of Social and Emotional Skills in Cities*" (EDU-CERI-CD(2015)13.en), by O. P. John & F. De Fruyt, 2015. OECD Publishing.

self-management skills, (b) using social awareness and interpersonal skills to establish and maintain positive relationships, and (c) demonstrating decision-making skills and responsible behaviors in personal, school, and community contexts (Durlak et al., 2011). The components of these goals—self-awareness, self-management, social awareness, relationship skills, and responsible decision-making—have become the major constructs around which CASEL's work is organized. The "CASEL 5" did not result from a scientific analysis of either SEL programs or individual differences, but instead informally emerged to characterize aspects and points of emphasis of various SEL programs, using language that policy makers found useful. SEL language has been adopted widely in government circles, such as in state and national standards, television programs, and other media.

CASEL maintains an extensive website with resources and links. CASEL focuses on SEL programs rather than assessments, but hosted an assessment work group for several years, resulting in several useful reports (Assessment Work Group, 2019) and resources, including the SEL Assessment Guide (CASEL, 2020b), along with a blog, a comparison of frameworks (Blyth et al., 2019), an alignment tool, webinars, and assessments emerging from several years of design challenges. CASEL has produced summaries and meta-analyses on the efficacy of SEL programs (Durlak et al., 2010, 2011; Mahoney et al., 2018; Taylor et al., 2017), a handbook on research and practice (Durlak et al., 2015); and position papers based on research findings and working with stakeholders on the CASEL website.

National Academy of Sciences/National Research Council

The National Research Council (2008, 2011, 2012) and its renamed successor, the National Academies of Sciences, Engineering, and Medicine (2017), conducted a series of workshops and consensus reports on interpersonal and intrapersonal skills. The 2012 consensus report reviewed several 21st-century skills frameworks and created a synthesized framework, organized around models from individual differences studies in abilities (Carroll, 1993) and personality (John & Srivastava, 1999). The taxonomy posited a set of cognitive (cognitive processes and strategies, knowledge, creativity), interpersonal (teamwork and collaboration, leadership), and intrapersonal (intellectual openness, work ethic, positive core self-evaluation) competency clusters. Each competency cluster (e.g., teamwork and collaboration) was in turn cross-walked to 21st-century terminology taken from the source documents (communication, collaboration, teamwork, cooperation, coordination, interpersonal skills, empathy, perspective taking, trust, service orientation, conflict resolution, negotiation), an O*NET descriptor or skill (Peterson et al., 1999; social skills), and a basic factor from the individual differences literature (e.g., Agreeableness). The value of organizing around existing models of abilities and personality is that such models provide an empirical foundation for categorizing constructs, in much the same way that John and De Fruyt's (2015) analysis does (see Table 19.1). This is a way to address the problem of jingle (a single term to describe multiple constructs) and jangle (multiple terms to describe the same

construct; Reeves & Vanator, 2014). It also provides a way to categorize interventions, with an expectation that specific construct interventions will provide spillover effects on within-category constructs in accord with a transfer gradient.

The National Academies (2017, p. 2) also convened a group to

> examine how to assess interpersonal and intrapersonal competencies (e.g., teamwork, communication skills, academic mindset, and grit) of undergraduate students for different purposes . . . to include identifying a range of competencies that may be related to postsecondary persistence and success, and that evidence indicates can be enhanced through intervention. (p. 2)

They identified eight competencies: (a) "behaviors related to conscientiousness . . . to self-control, hard work, persistence, and achievement orientation"; (b) "sense of belonging—a student's sense that he or she belongs at a college, fits in well, and is socially integrated"; (c) "academic self-efficacy—a student's belief that he or she can succeed in academic tasks"; (d) "growth mindset—a student's belief that his or her own intelligence is not a fixed entity, but a malleable quality that can grow and improve"; (e) "utility goals and values—personal goals and values that a student perceives to be directly linked to the achievement of a future, desired end"; (f) "intrinsic goals and interest—personal goals that a student experiences as rewarding in and of themselves, linked to strong interest"; (g) "prosocial goals and values—the desire to promote the well-being or development of other people or of domains that transcend the self"; and (h) "positive future self—a positive image or personal narrative constructed by a student to represent what kind of person he or she will be in the future" (pp. 5–6). The report stated that "self-report methods, with their known limitations, predominated in the assessments of the eight competencies" (p. 8) and that "analysis of the quality of the assessments used in the intervention studies revealed spotty attention to reliability and almost no reported evidence of validity or fairness." (p. 8).

OECD's SSES framework

OECD manages the PISA (OECD, 2019b) and the Program for the Assessment of Adult Competencies (PIAAC), among others. PISA and PIAAC are surveys measuring respondents' cognitive proficiency and include background questionnaires. OECD recently launched the Study on Social and Emotional Skills, an international cross-sectional (ages 10 and 15) survey of students, teachers, and parents focusing on social and emotional skills, conducted in 12 city sites (Kankaraš, 2017; OECD, 2019a). The long-term goal is to monitor the growth of social-emotional skills from school to the labor force in a policy-relevant, feasible, valid, reliable, comparable, ethical, cost-effective, and sustainable way.

The project has published several framework documents (Chernyshenko et al., 2018; John & De Fruyt, 2015; Kankaraš & Suarez-Alvarez, 2019; OECD, 2015). OECD (2015) provided the justification for the study with a comprehensive literature review, concluding that social and emotional skills, as important as they are, were hard to measure, and consequently teachers, schools, and policy makers

do not know "if their efforts at developing these skills are paying off," and that these skills "are seldom taken into account in school and university admissions decisions" (p. 13).

OECD (2019a) proposed the Big Five as an organizing framework because of positive research on the psychometric qualities of Big Five measures (reliability, predictions of outcomes) and a sound justification for five independent dimensions. Even with only five dimensions, the model is comprehensive, particularly when lower order correlated dimensions are added (Dudley et al., 2006; John & Srivastava, 1999; Paunonen & Ashton, 2001). Also, the Big Five has a scientific underpinning in the lexical hypothesis (Goldberg, 1993), that language evolves to accommodate the most important and salient differences between people in the form of descriptors. There are competitors to the Big Five, such as the six-dimension hypothesis (Lee & Ashton, 2004), but this a relatively minor variation. Major alternative explanatory models are not readily available. Seeming alternatives, such as 21st-century skills, can easily be shown to be largely captured by the Big Five, as John and De Fruyt's (2015) analysis showed (see Table 19.1).

The OECD (2019a) framework proposed five domains with several facets within each domain: task performance (achievement motivation, responsibility, persistence, self-control), emotional regulation (stress resistance, optimism, emotional control), collaboration (empathy, trust, cooperation), open-mindedness (tolerance, creativity, curiosity), and engaging with others (sociability, assertiveness, energy). These are the Big Five factors (Conscientiousness, Neuroticism, Agreeableness, Openness/Intellect, and Extraversion, respectively) with more palatable names. The framework adds a sixth dimension, compound skills, which includes critical thinking, metacognition, and self-efficacy. Unlike some of the frameworks reviewed here, this framework is not merely notional or a helpful organizing scheme, but a structural model and, with the assessments being administered, a measurement model, which can be evaluated with confirmatory factor analysis. A report on preliminary findings is available (OECD, 2021).

UC Consortium on Chicago School Research (UCCCSR)

UCCCSR conducted a comprehensive literature review to identify the noncognitive factors critical to children's learning and development (Farrington et al., 2012). They sought to identify factors that were linked to academic achievement, at all ages and academic levels, that were simultaneously "not fixed traits," but ones that could be shaped by the educational context. The nature–nurture distinction is fuzzy and often overblown, but a perception of fixed traits versus malleable competencies has influenced the research conducted in this area. UCCCSR's literature review can be viewed as a framework for noncognitive factors. They proposed five categories of factors: (a) academic behaviors (going to class, doing homework, organizing materials, participating, and studying); (b) perseverance (grit, tenacity, delayed gratification, self-discipline, self-control); (c) mindsets (sense of belonging, growth mindset, self-efficacy, belief in the value of academic work); (d) learning strategies (study skills, metacognitive

strategies, self-regulated learning, goal setting); and (e) social skills (interpersonal skills, empathy, cooperation). There has been some empirical research evaluating this framework (Farruggia et al., 2016; Wanzer et al. 2019).

Other Frameworks

The range of noncognitive variables goes beyond the Big Five and the constructs identified thus far. Here, we discuss proposals that expand the list of potentially important constructs.

Social Attitudes

Saucier (2000, 2013) identified several social attitude dimensions (or belief system components: beliefs, attitudes, ideologies, worldviews) using a lexical approach. Respondents were given 389 descriptions of terms ending in "ism," such as authoritarianism, liberalism, conservatism, and communism, sampled from the dictionary. They rated their agreement with the descriptions, and factor analyses resulted in four factors—(a) alpha (tradition-oriented religiousness, e.g., creationism); (b) beta (unmitigated self-interest, e.g., hedonism); (c) gamma (communal rationalism, e.g., utilitarianism); and (d) delta (subjective spirituality, e.g., reincarnationism). Saucier (2013) included additional items identifying a fifth factor, inequality-aversion (e.g., egalitarianism vs. elitism, jingoism). He found traditional religiousness correlated with Republican Party (within the United States) preference. Inequality aversion, subjective spirituality, and communal rationalism correlated with Democratic Party preference; and unmitigated self-interest correlated negatively with Openness/Intellect and Agreeableness.

Economic Preference Parameters

Almlund et al. (2011, Table 6), proposed a set of economic preference parameters—time, risk, and social preferences—based on the behavioral economics literature and argued that they overlap Big Five factors. Time preference (delay discounting) relates to one's ability to delay gratification; risk preference refers to the amount of risk one is willing to assume; and social preferences have to do with leisure, altruism, trust, and positive and negative reciprocity. There are conceptual overlaps—risk-taking is a facet of Extraversion, time preference (punctuality vs. procrastination) is related to Conscientiousness, and trust is related to Agreeableness. A. Falk et al. (2015, 2016) administered measures of economic preferences along with the Big Five and found some overlaps; they also found correlations with national indicators such as educational attainment, savings rate, and risk behaviors (e.g., smoking).

Kyllonen (2016) proposed (but did not test) a taxonomy to summarize all these constructs. The major categories were the Big Five, generalized attitude dispositions, interests, personal beliefs, cultural and behavioral norms, social axioms, attitudes toward school, ivalues, subjective well-being, economic preferences (time, risk, and social preferences), emotional intelligence, metacognition, creativity, collaboration, cognitive bias susceptibility, emotions, moods, and states of mind.

National Assessment of Educational Progress

Several other frameworks should be mentioned because of their importance in the field. One is for background (contextual) questionnaires for large-scale assessments, such as the National Assessment of Educational Progress (NAEP) (National Assessment Governing Board, 2013), PISA (OECD, 2017b), and other major assessments, such as TIMSS, PIRLS, and PIAAC. The National Assessment Governing Board (2013) identified priorities for NAEP contextual data collection, which were (a) NAEP reporting categories (socioeconomic status, gender, race/ethnicity, disability status, and English language status); (b) contextual factors with relationships to achievement; and (c) subject-specific information, with priority based on validity, reliability, universality (can be collected from all students), currency, respondent burden, logistic feasibility, cost-effectiveness, timeliness, nonintrusiveness, whether trends are important, and whether the information contained is valuable for understanding academic performance and how to improve it. NAEP is administered every year, and there are many different questionnaires (student, teacher, school), which are changed regularly. The student questionnaire is limited to 15 minutes. The National Center for Education Statistics (n.d.-b) provides details and the questionnaires themselves.

PISA

PISA measures reading, mathematics, and science every 3 years (since 2000; highlighting a different subject each cycle) to about half a million 15-year-olds in 70+ countries and includes a 30-minute student questionnaire. The PISA 2018 questionnaire framework (OECD, 2019b), highlighting reading, includes three construct categories: student background (e.g., out-of-school reading, socioeconomic status), schooling (teacher qualifications, instruction time, parental involvement), and noncognitive/metacognitive. This latter category includes attitudes motivation, strategies dispositional variables (achievement motives; incremental mindset; perseverance; subjective well-being; information and communications technology, motivation, and practices) school-focused variables (learning beliefs, attitudes toward school, achievement goals), and dispositions for global competence (included because global competence was assessed in the PISA 2018 survey). The latter includes communication and relationship management, knowledge of and interest in global developments, Openness and flexibility, and emotional strength and resilience. There are common themes and scales across the PISA cycles as well as changes (see OECD, 2013a, 2016, 2019b, for the past three questionnaire frameworks). There have been many analyses of the constructs together to determine the dimensionality and structure of these scales (J. Lee & Stankov, 2013; Marsh et al., 2006).

The California Office to Reform Education (CORE)

CORE is a coalition of eight California school districts serving over half a million students. It was the first U.S. jurisdiction to obtain a waiver from the U.S. Department of Education's No Child Left Behind legislation to administer a noncognitive assessment for the purposes of accountability. The questionnaire measured four constructs in

students—self-management, growth mindset, self-efficacy, and social awareness (West, 2016). There was not a systematic process for identifying these constructs, but they are some of the more popular in K–12 discussions. Two come from the CASEL five; growth mindset comes from Dweck (2017), but is related to locus of control and attribution theory, which have been a part of education discussions for decades (Graham, 1991; Rotter, 1966), as has self-efficacy (Bandura & Schunk, 1981; Schunk, 1989).

Mission Statements

Oswald et al. (2004) identified 12 critical skills for college by sorting statements coming from college mission statements. They clustered into intellectual (knowledge, continuous learning, and curiosity/artistic appreciation), interpersonal (multicultural tolerance, leadership, interpersonal, social responsibility), and intrapersonal (health, career orientation, adaptability, perseverance, ethics) behaviors. With a similar methodology, Stemler et al. (2011) identified cognitive, social, emotional, and civic as the most common themes in high school mission statements.

Competency Identification

Shultz and Zedeck (2011) conducted extensive interviews with practicing lawyers, asking questions such as, "If you were looking for a lawyer for an important matter for yourself, what qualities would you most look for?" They identified 26 effectiveness factors, divided into eight categories (cognitive, research and information gathering, communications, planning and organizing, conflict resolution, entrepreneurship, working with others, character). This approach has also been used in developing competency frameworks, which are widely used in business for promoting assessment and training needs. One example is SHL's Great Eight (Bartram, 2005). These eight high-level competencies have been mapped to the Big Five (the letter in parentheses indicates the competency's highest correlate (A = Agreeableness, E = Extraversion, O = Openness; N = Neuroticism; C = Conscientiousness; g = cognitive ability): learning and deciding (E); supporting and cooperating (A); interacting and presenting (E); analyzing and interpreting (g); creating and conceptualizing (O); organizing and executing (C); adapting and coping (–N); enterprising and performing (–A). Each of these is divided further into specific competencies. Leading and deciding is divided into deciding and initiating action and leading and supervising. Deciding and initiating action is further divided into six competencies, including making decisions, acting on own initiative, and taking calculated risks. Other competency frameworks are Clifton Strengths (Rath & Conchie, 2008) and Lominger's 67 competencies (Lombardo & Eichinger, 2004), which include competencies such as dealing with ambiguity, creativity, motivating others, planning, and building effective teams. Competencies are typically identified based on factor analysis of responses to rating scale items written to represent a broad variety of attitudes, values, and behaviors relevant to workplace performance, providing an empirical foundation to this work. But there is less emphasis on finding commonalities across competency frameworks (because of market incentives) compared to the case with personality psychology.

Personality Testing Frameworks: The Big Five and Facets

Personality assessments are built around frameworks. The Big Five is an important framework. However, the most popular personality assessments typically provide many more than 5 dimensions. Analyses of items from the International Personality Item Pool (Goldberg et al., 2006), a repository of over 3,000 personality items aligned with over 400 commercial instrument scales, reveal over 25 factors (Condon, 2018). Major commercial instruments measure more than 5 dimensions: SHL's OPQ (32 dimensions); Gallup's StrengthsFinder 2.0 (Rath, 2007) (34 dimensions, called themes), and Lominger (Lombardo & Eichinger, 2004) (67 competencies). A. Costa and Kallick (2008) proposed 16 habits of mind, essential characteristics for success, including persisting, managing impulsivity, and thinking independently. C. Peterson and Seligman (2004) proposed 26 character strengths rolled up into 6 classes of virtues. A consistent finding is that the 5-factor model (or the 6-factor HEXACO model, Lee & Ashton, 2004) is the highest order of a hierarchical model (John & Srivastava, 1999; Drasgow et al., 2012). Although facets (lower order components) tend to be less stable across samples and languages than the Big Five, the 5- (or 6-) factor model is the best model we have of the major independent dimensions of human descriptions of personality, with additional structure of maybe 20 or 30 reliable factors beneath the Big Five.

Comparison of Frameworks

The frameworks reviewed only scratch the surface—Berg et al. (2017) identified 136 frameworks in the K–12 area alone! And there are higher education frameworks, such as ACT's holistic framework (Camara et al., 2015). A question is, How can such frameworks be evaluated or compared? A report from the CASEL assessment work group (Blyth et al., 2019) proposed 10 criteria. Five concerned conceptual clarity: specificity (defines competencies), balance (includes interpersonal, intrapersonal, and cognitive competencies), developmental (specifies how development occurs), culturally sensitive (accommodates cultural differences), and empirically grounded (reflects findings of associations between competencies and school, work, and life success). Five concerned implemental support: intended for practice, resources for practitioners, resources for use by children and youth, resources for measurement and data use, and empirically tested. The criteria were developed with practitioners in mind—school staff seeking a framework for implementation in response to district, state, or federal policies or mandates. Related CASEL efforts include a white paper on what frameworks are and why they are useful (Blyth et al., 2018) and tools for selecting and cross-walking frameworks (Jones et al., 2019).

The Big Five personality model has been influential in framework development. OECD's Study on Social and Emotional Skills program has adopted it, and workforce testing has embraced it. Long-standing workforce tests are issuing five-factor model conversions and score reports, and new assessments are specifically tailored around the Big Five. Why is that? The five-factor model likely has more research—reliability, prediction studies, and, increasingly, personality change studies—than any other model of individual differences in social and emotional dimensions. The Big Five has become a

standard and therefore a means to align findings and advance the field quickly. CASEL language is more popular, certainly within K–12 education, and it is highly represented in policy statements and in state social-emotional learning standards (SEL itself is a CASEL coinage). The Big Five's power is that it emerged from an empirical finding regarding the words we use to describe ourselves and others (Goldberg, 1993). For these reasons, the Big Five model and CASEL's five competencies are the dominant frameworks today. Few, if any, frameworks can satisfy all the requirements laid out by Blyth et al. (2019). It is likely that for the foreseeable future no one framework will be adopted universally for diverse uses. As frameworks are implemented, for various purposes, and results are analyzed, we may see more consolidation based on what kinds of frameworks prove most useful.

METHODS

Separating methods and constructs is difficult. The construct of personality is almost completely confounded with the rating scale method for measuring it. In the research literature and in policy discussions, mention of personality is invariably linked with rating scale measurement. We suspect that resistance to implementing personality assessment in schools has to do with the limitations of rating scale methodology. Despite widespread agreement on construct importance, there may be doubt about the adequacy of ratings scales for measuring them. Imagine a world in which, instead of mathematics tests, we administered self-ratings of mathematics knowledge and ability. Significant decisions would not likely be based on such data.

Disentangling method and construct is useful conceptually and for the field to advance, which has been acknowledged for over half a century (Campbell & Fiske, 1959). There may be methods other than rating scales for measuring personality. *Persistence* can be measured variously:

- Self-rating: "How often do you work on a task until you are finished: Rarely/sometimes/often/always or almost always?"
- Teacher rating: "X completes their assignments on time: True/Not true?"
- School administrative record: Missed assignments____; tardiness____
- Performance test: Length of time spent adding columns of numbers before requesting a break
- Situational judgment test: "You have a test the next day and don't feel fully prepared. You are very tired and you are not thinking clearly. What do you do?"
- Behavioral interview: "Tell me about a time when you had to persist on a task despite many barriers in your way."

Here, we review the main methods used to measure intra- and interpersonal skills. We review the main issues in the design, development, and scaling of assessments and the key evidence and challenges with respect to reliability, validity, and fairness (for additional information on these three concepts, see Lee & Harris; Lane & Marion; Zwick; and Rodriguez & Thurlow, all in this volume).

Self-Report Ratings

Self-report ratings are rating scale measures, known as Likert-scale measures (Likert, 1932). They refer to a statement or a question a respondent is asked to express agreement or other judgment about (e.g., behavior frequency). An example is, "How much do you agree with the statement, I am a hard worker? Strongly agree/agree/neither agree nor disagree/disagree/strongly disagree," or "How often do you come to class prepared? Never or almost never/seldom/sometimes/often/always or almost always." The research we have reviewed to this point is based on this method. The contextual questionnaires in NAEP, PISA, PIAAC, World Health Organization's Health Behavior in School-Aged Children survey (Inchley et al., 2018), and most assessments in RAND's Education Assessment Finder (Hamilton et al., 2018) and CASEL's (2020b) SEL Assessment Guide are based on this method, as are the assessments used in the CORE Districts and the items in ETS's SuccessNavigator. The Big Five personality model is based on self-report ratings.

There are general rules for item writing (Alreck & Settle, 1994; Fowler, 2006; National Assessment Governing Board, 2002). Items should be precise (e.g., regarding who the question refers to, the time frame), concise (e.g., ask one question, one topic; avoid "double-barreled" questions; do not use more words than necessary), neutral (e.g., avoid leading the respondent to a preferred answer), and simple (avoid esoteric vocabulary or complex sentence constructions). An easier rule of thumb is the headline rule. Imagine a *New York Times* headline that summarizes findings from the survey based on one item. "40% of New York 4th graders agree that what they learn in school is important for their future" is fairly interpretable. "30% of New York 4th graders *often believe* that what they learn is irrelevant" is probably not—what does "often believe" mean?

The NAEP contextual questionnaire group (Bertling, 2015) developed a format taxonomy for rating scale questions, applicable to questionnaire assessments broadly. Their 12-scale taxonomy specified *frequency* ("how often") versus *amount/extent* ("how much"); frequency scales were abstract or quantified, with abstract being *absolute* (without reference point) or *relative* (with a reference point). Two types of quantified scales are absolute (i.e., counts) and *average quantified frequencies* (average counts within a time period). There are eight amount/extent scale types: time, emphasis, possibility to change, similarity to self ("like me"), agreement, confidence, likelihood, and importance. Each is associated with specific recommended response categories. An absolute abstract frequency scale uses the categories "never or hardly ever," "once in a while," "sometimes," "often," and "always or almost always." An average quantified frequency scale uses the categories "never," "about once or twice a year," "about once or twice a month," "about once or twice a week," "every day or almost every day," and "several times a day." An agreement scale uses the labels described in the previous paragraph; a similarity-to-self scale uses "not at all like me," "a little bit like me," "somewhat like me," "quite a bit like me," and "exactly like me." This was implemented in PISA 2022 (OECD, 2023).

Research issues include how many response categories to use (e.g., 2, "true of me" versus "not true of me" to 11 or more, such as a continuous graphic scale) (Revilla et al., 2014,

suggested 5 for agreement scales) and whether to include a neutral middle point (e.g., "neither agree nor disagree") (Moors, 2008, suggested probably not). There have been studies on the interpretation of frequency quantifiers, such as seldom and often (Bocklisch et al., 2012; Newstead & Collis, 1987), which may be useful in making response category labels by spreading out the categories (selecting category labels corresponding to a spread of specific percentages that might be informative). But determining the most effective approach for a particular context is difficult. An alternative is analyzing response data after data collection to determine category frequencies and to explore the relationship between item responses and the underlying trait level using nonparametric methods, such as nonparametric option characteristic curves (Ramsay, 1991), or item response theory (IRT) methods, such as the nominal response model (Bock, 1972). R. H. Meyer et al. (2018) provided examples with the California CORE data.

Reliability

Two general approaches to analyzing and scoring rating scale data are classical test theory (CTT) and item-response theory (IRT). The CTT approach models test (or scale) scores (with exceptions; Legree et al., 2021). It assumes an observed test score is the sum of a true score and a random error. Ratings are assumed to be on an interval (continuous) response scale. A rule of thumb is that more than six ordered response category responses can be treated as continuous (Millsap, 2011, p. 121), but with fewer, they should be treated as categorical (Lubke & Muthen, 2004). Nevertheless, treating rating scale data as continuous is still the most common approach. The method involves transforming category responses (e.g., "strongly agree," "agree," "disagree," "strongly disagree") to numbers, usually 1 to k (sometimes, 0 to $k - 1$), where k is the number of categories, and then computing descriptive statistics (mean, standard deviation, skewness), and reliability (usually Cronbach's alpha) from the continuous variable. West et al. (2020) used a CTT approach when analyzing the CORE Districts data. A self-management scale (e.g., "I came to class prepared") used a five-category response scale ("almost never," "once in a while," "sometimes," "often," and "almost all the time"). The scale metric was transformed from the five scale categories, yielding a mean of 4.05 (presumably, the nine self-management items mean), a standard deviation of .69 and an alpha of .85 was reported (their Table 2). All four construct scales (self-management, growth mindset, self-efficacy, and social awareness) had five response scales, and although they had different labels ("not at all true" to "completely true" for growth mindset; "not at all confident" to "completely confident" for self-efficacy; "not carefully at all" to "extremely carefully" for social awareness), this allowed some degree of comparability across scales.

CLASSICAL TEST THEORY RELIABILITY In CTT, reliability is the proportion of true score to observed score variance, $r_{xx'} = \sigma_t^2 / \sigma_x^2$. Reliability is a function of both the average interitem correlation (or covariance) and the number of items in the scale, as can be seen from a formula for alpha, $r_{xx'} = \dfrac{n\overline{c}}{\overline{v} + (n-1)\overline{c}}$ where n is the number of items

in the scale, $\bar{c}$ is the mean interitem covariance, and $\bar{v}$ is the average item variance.[3] One implication of this is that scales can be made more reliable simply by adding more items. This is shown in the Spearman–Brown prophesy formula, $r^*_{xx'} = \frac{kr_{xx'}}{1+(k-1)r_{xx'}}$, where $r^*_{xx'}$ is the expected new reliability resulting from changing the test length by a factor of k ($k = 2$ to double the number of items or $k = .5$ to halve the number).

Reliability is important when correcting correlations with external variables and when determining effects of changes in reliability on changes in correlations with external variables. Disattenuated correlations between a scale, x, and an external variable, y, that is, correlations corrected for unreliability in both x and y, are computed as $r^*_{xy} = r_{xy}/\sqrt{r_{xx'}\, r_{yy'}}$. This can also be used for estimating changes in correlation with external variables, from r_{xy} to r_{xy2} based on changes in reliability, from $r_{xx'}$ to $r_{xx'2}$ is $r_{xy2} = r_{xy}\sqrt{r_{xx'2}/r_{xx'}}$; that is, the new predictive correlation is increased by a factor of the ratio of the square roots of the new to old reliabilities (Nunnally & Bernstein, 1994, eq. 7-1, p. 257).

IRT SCORING AND RELIABILITY IRT can be used to analyze and score data from rating scales. IRT treats the ratings as categorical variables and models the probability of choosing each rating category as a function of item parameters and person latent trait. There are many IRT models—van der Linden (2016) comprises 33 chapters, each dedicated to an IRT model (or class of models).[4]

There are ways to taxonomize IRT models to focus on the ones most important for the analysis of rating scales. First, distinguish continuous from categorical latent variables. Categorical latent variable models include mixture models and latent class models. These are relatively rare in analyses of rating scale data, although there are some applications for detecting response styles and faking (Eid & Zickar, 2007).

The most common IRT models assume a continuous latent variable underlying the ordered categorical responses. What is being modeled is the probability of selecting one of the k categories, such as "agree" or "strongly agree." A set of "divide-by-total" polytomous IRT models (Thissen & Steinberg, 1986) that share common parameterization but with decreasing degree of constraints includes rating scale models (RSM; Andrich, 2016), partial credit models (PCM; Masters, 2016), generalized partial credit models (GPCM; Muraki & Muraki, 2016), and nominal response models (NRM; Bock, 1972). Item parameters for these models include item/category slope and category intercept parameters, which, respectively, represent the discrimination of the item/category as a function of the latent trait and the relative endorsement of a category.

The RSM (Andrich, 2016) assumes all item slopes are the same, and the differences of category intercepts between adjacent categories (e.g., disagree and agree, agree and strongly agree) are the same across items. The PCM assumes equal item slopes, but relaxes the constraints on category intercepts. The GPCM further relaxes the assumption of equal item slopes for all items in the scale, an assumption that characterizes

both the RSM and the PCM. However, discrimination is common across all categories. The NRM even further allows different categories to have different slopes. These are IRT models ordered by flexibility. Fitting these models to the same data can be used to test out different model assumptions (Preston et al., 2011). A cost for flexibility is the requirement for larger sample sizes. De Ayala (2009) recommended minimum sample sizes of 250, 250, 500, and 600 for the RSM, PCM, GPCM, and NRM, respectively. Once an IRT model has been determined, an individual can be scored as the estimate of the latent trait in the given IRT model.

In IRT, test information and measurement error vary with the latent trait. To get overall reliability averaging across the latent trait, Green et al. (1984) suggested marginal reliability. Following the CTT reliability definition of variance of true scores over variance of observed scores, the IRT marginal reliability is calculated as $(\sigma_\theta^2 - \bar{\sigma}_e^2)/\sigma_\theta^2$, where σ_θ^2 is the variance of latent trait and $\bar{\sigma}_e^2$ is the average error variance across the latent trait distribution.

CTT VERSUS IRT FOR RATING SCALE METHODS Most personality research and noncognitive assessment research more generally employs CTT methods. CTT methods are convenient; well known; easy to apply with a wide variety of software, including spreadsheet software; and familiar to researchers and stakeholders; they can also be appropriately applied in small samples. IRT models are more complex, less well known, require specialized software, and generally require larger samples. But there are significant advantages to IRT methods. IRT overcomes the fundamental item–person confounding problem in CTT, separating item and person effects (van der Linden, 2016). This separation leads to many advantages: IRT can be used to provide better item diagnostics, scoring of test takers based on response patterns, form assembly, equating and linking of forms, assessment of change, and examination of measurement invariance across test-taking subgroups. These advantages are routinely discussed in IRT textbooks (Reise & Revicki, 2015), and the field of educational measurement has embraced IRT approaches. Because educational measurement increasingly includes personality and noncognitive assessment, a similar adoption of IRT methods in this realm can be expected.

Validity

EVIDENCE BASED ON CONTENT Rating scale items are simply statements about which a respondent expresses agreement (or frequency, amount, or extent). So, if respondents indicate that they "strongly agree" that "I am a hard worker," this can be taken at face value as an indication of Conscientiousness, by the definition of Conscientiousness. Messick (1995) referred to this as content relevance. This prima facie case is bolstered by additional evidence that ratings on other indicators of Conscientiousness (e.g., "I am thorough in my work," "I complete tasks in an efficient manner") tend to intercorrelate. The literature on item intercorrelations based on the five-factor model attests to the nature of those relationships (John & Srivastava, 1999).

This interpretation can be challenged based on *calibration* and *faking* (or social desirability responding). The calibration challenge is that interpreting Respondent A's "strongly agree" that "I am a hard worker" as an indication of a higher level of Conscientiousness than Respondent B's "agree" that "I am a hard worker" (aside from the mapping of "hard worker" to "Conscientiousness," discussed in the previous paragraph) assumes that the two respondents have similar understandings of both the stem and the categories (Böckenholt, 2004). An example of miscalibration can be found in the meta-analytic correlation between self-estimates of cognitive ability and cognitive ability measured by tests: they are only correlated around $r = .33$ (Freund & Kasten, 2012).

The faking challenge has to do with respondents presenting themselves in a too-favorable light (self-presentation); social-desirability bias (Paulhus, 2002). Niessen et al. (2017) demonstrated self-presentation with university applicants who completed a set of ratings and then repeated the ratings after being selected into the program (reduced self-presentation). Scores were higher and outcome predictions were lower with self-presentation. West et al. (2018) minimized self-presentation by assuring anonymity, attaching no stakes to the assessment, and having the proctor (teacher) stay in the back of the room rather than roam the aisles looking over students' shoulders. Still, self-presentation presents a challenge, particularly in a high-stakes context.

EVIDENCE BASED ON CORRELATIONS WITH OTHER MEASURES The other key source of validity evidence for rating scale measures is in their correlations with external measures. Meta-analyses have shown moderate correlations between personality self-ratings and grades in higher education (McAbee & Oswald, 2013; Noftle & Robins, 2007; Poropat, 2009; Richardson et al., 2012; Robbins et al., 2004, 2009; Vedel, 2014). Conscientiousness is the highest correlate, typically in the mid .20s range. These studies, as well as the ones cited next, are typically conducted in low-stakes research settings, giving respondents little incentive to fake self-ratings.

West et al. (2018) found moderate correlations within grade between the four CORE Districts scales (self-management, growth mindset, self-efficacy, and social awareness) and grade point average (GPA; $r = .18$ to $.44$) and math ($r = .14$ to $.37$) and English language arts test scores ($r = .15$ to $.34$) (*N*s ranged from 7,893 to 127,134). This aligns with expectations based on our review and supports a test score interpretation that the four constructs are associated with educational success.

There are two key challenges to this interpretation, independent of the challenges associated with the calibration and faking issues discussed previously. The four scores are correlated quite highly, relative to their reliability. At the school level, correlations range from $r = .36$ to $r = .97$, with a median $r = .91$ (Hough et al., 2017). This suggests that there might not be four unique pieces of information in the four scale scores, which would challenge an interpretation based on the scale content (e.g., self-efficacy, growth mindset), perhaps toward an alternate interpretation based on the common elements across the scales (e.g., social desirability). The fact that West (2016) showed that the highest correlate of English language arts grades was the composite of all four measures,

rather than any individual measure, supports this idea. A subscore analysis, as suggested by Haberman (2008), would be informative to establish scale independence.

Second, interpreting results across grades is challenging. Figure 19.1 shows the cross sectional comparisons of CORE District students' SEL scores from Grades 4 to 12 and, for comparison purposes, the math and English language arts achievement test scores. Achievement test scores, which are vertically scaled and therefore comparable, are monotone, almost linear increasing. The SEL scores are from identical questions (i.e., the same questions were asked in the 3rd through 12th grades), scaled to a common metric. They sometimes go up, sometimes down, without a clear pattern. At the least, this challenges the interpretation of test scores as indicating an association of the four constructs with achievement. Fourth graders' social awareness is considerably higher than that of high schoolers, yet their achievement is considerably lower. Within a grade, the correlation is higher. The cross-grade pattern is difficult to interpret. Generally, age, period, and cohort effects all might influence findings like these (Ion et al., 2022).

Fairness

Fairness in assessments as well as communications about those assessments, in various standards documents (American Educational Research Association [AERA] et al., 2014; ETS, 2014, 2022), refers to the requirement that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results or their interpretation. Fairness reviews conducted prior to pilot testing consider representation of the groups to be studied, accommodations, presentation and response formats, the potential effect of timing (e.g., time limits), language (e.g., English language learning status), experience with the technology, and prohibition of inappropriate language, among other factors. Noncognitive and cognitive assessments are not different in these regards. Generally, the goal is to avoid materials that are not fair because they include the wrong content and skills or they fail to include a good sample of the right content and skills (ETS, 2022).

MEASUREMENT INVARIANCE Measurement invariance analyses is conducted to ensure comparability of interpretation across groups (e.g., race/ethnicity, gender, age, language) or over time. With continuous data there are well-documented approaches to establishing measurement invariance through a sequence of tests for configural, metric, and scalar invariance. Tests for configural invariance establish that the factor structure is the same across groups; metric invariance establishes that item loadings are the same; scalar invariance establishes that the item intercepts and loadings are the same. Establishing metric invariance allows one to interpret the correlations between scale scores and external variables as comparable for the different groups; establishing scalar invariance allows one to interpret the mean scale (factor) differences between groups. Failure to establish measurement invariance at any point makes comparisons between groups problematic. It is not appropriate to claim a mean "Conscientiousness" difference between males and females if scalar invariance is not established. If a Conscientiousness item was "I enjoy assembling and disassembling car engines" and the mean difference on

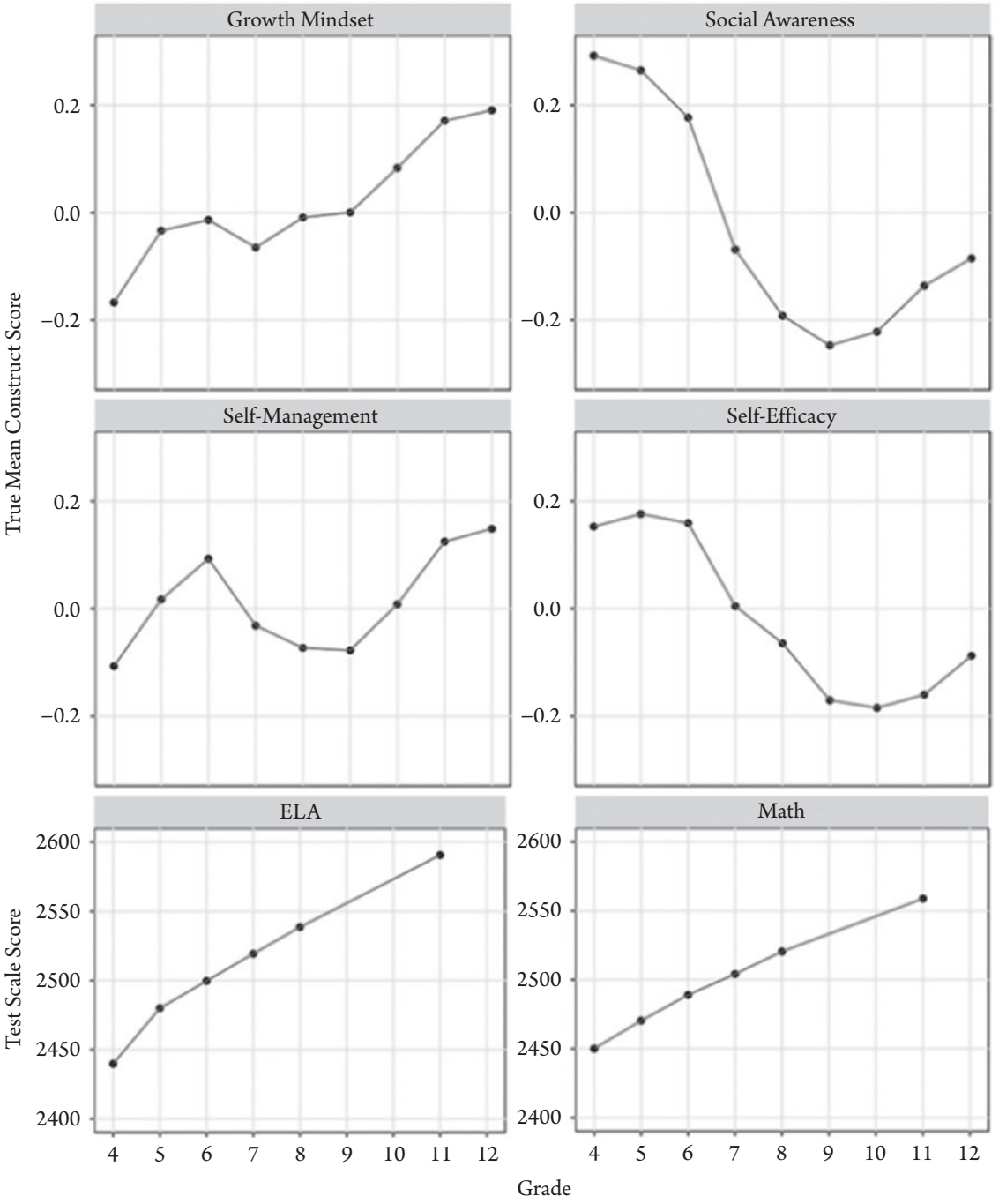

FIGURE 19.1

Cross-Sectional Comparisons of CORE District Students Scale Scores Across Social-Emotional Learning (SEL) Scales (Top Four Panels) and Achievement Test Scores (Bottom Two Panels)

Note. Achievement test scores are vertically scaled and monotone increasing. SEL scale scores are from identical questions across years scaled to a common metric. They are not monotone increasing. From *Trends in Student Social-Emotional Learning: Evidence from the CORE Districts* (Working Paper), by M. R. West, L. Pier, H. Fricke, H. Hough, S. Loeb, R. H. Meyer, and A. B. Rice, 2018, PACE: Policy Analysis for California Education, CORE-PACE Research Partnership. Reprinted with permission.

a Conscientiousness scale that included the item showed a higher mean for males, one might be reluctant to conclude that males were more conscientious than females (Ployhart & Oswald, 2004).

With categorical data, as typically found on rating scales, the configural–metric–scalar series of tests is not appropriate, unless those data are treated as continuous. But if they are treated as categorical, an alternative sequence of first testing for configural invariance, then for threshold invariance, then for loadings and intercept invariance is recommended (Svetina et al., 2020; Wu & Estabrook, 2016). There are several reasons for conducting measurement invariance analyses. One is to determine the appropriateness of comparing groups. Another is to explore reasons for misfit, which can provide additional insight into group differences (Putnick & Bornstein, 2016). Typically, when conducting measurement invariance tests, steps are taken to identify misfitting items and delete them or relax certain assumptions about them (e.g., thresholds). This process can provide insights on possible reasons for misfit and allow for comparison of scale scores between groups. The practical importance of lack of invariance should also be considered (Nye & Drasgow, 2011; Schmitt & Ali, 2015).

Response Style

For rating scale items, response style is potentially a source of construct-irrelevant variance. Response style is a major factor in international assessments involving different languages, but it also can also differentiate subgroups within a language group (e.g., gender, race/ethnicity, immigrant status). There are well-documented differences between cultural groups in using the response scale, independent of the construct being measured. These include extreme response style (tendency to choose scale extremes, such as "strongly disagree" or "strongly agree"), modest or midpoint response style (tendency to choose the scale midpoint, such as "neutral" or "neither agree nor disagree"), and yeah-saying or acquiescence response style (ARS; tendency to agree). Response style effects can be adjusted for statistically. Khorramdel and von Davier (2014) identified extreme response style and modest or midpoint response style using a multinomial processing tree approach, showing the generalizability of response styles across response scales. He and van de Vijver (2015), using Teaching and Learning International Survey data, interpreted a general response style factor with positive loadings from extreme response style and socially desirable responding and negative loadings from modest or midpoint response style as response amplification versus moderation. Primi et al. (2019) identified positive and negative key item pairs (close to opposite) within a scale and presented both classical and IRT corrections to bias (typically positive bias, such as responding "strongly agree" to one of the pair and "neutral" or "disagree" to its opposite). C. F. Falk and Cai (2016) provided a general approach to modeling response styles. A special issue of the *British Journal of Mathematical and Statistical Psychology* (Khorramdel et al., 2019) assembled eight different approaches for correcting rating scale data for response style, including multidimensional IRT, IRTree, and clustering approaches. Some of these methods should routinely be applied to rating scale questionnaires.

Although rating scale approaches, particularly self-ratings, are by far the most common in the assessment of interpersonal and intrapersonal skills, there are significant challenges in their interpretation as construct measures. Also, while self-rating scales can be used in low-stakes assessments, they are problematic for high-stakes use (Niessen & Meijer, 2017).

Others' Report Ratings

Others' (informant) ratings refers to evaluations by others—peers, supervisors, professors, or others who are in position to evaluate the target. Unstandardized evaluations, such as letters of recommendation, are common, particularly in higher education and the workforce (personnel selection, promotions), but little research exists on them. Despite that, letters of recommendation and associated ratings are considered quite important in education admissions (Kuncel et al, 2014) and in the workforce (Judge & Higgins, 1998). Perhaps the reason for their popularity is that they protect against the faking and potential deception that can characterize self-reports. Personal statements are commonly required in admissions and scholarship competitions, despite little evidence for their value (Murphy et al., 2009). The reader of a recommendation letter can be at least somewhat assured that a relatively disinterested party is endorsing a candidate.

Research on letters of recommendation is based on post hoc numerical summaries of letter contents (Baxter et al., 1981; Kuncel et al., 2014). Such summaries correlate with outcome measures in higher education, GPA, and degree attainment and provide small, but incremental prediction beyond other measures in predicting attainment (Kuncel et al., 2014). But a quantitative summary of letter contents is not routinely available and is not how letters are typically used, which is more anecdotal. The future may bring natural language processing analyses of letters, as with essays and constructed response tasks (Madnani & Cahill, 2018), but little of that has appeared so far.

Many formats exist for others' ratings besides open-ended letters of recommendation. They range from a standardized letter of recommendation used for graduate admissions, such as ETS's Personal Potential Index (Kyllonen, 2006), to a clinical psychologist's rating of candidates' psychological fitness for military duty based on a 30-minute interview (Lindqvist & Vestman, 2011), a teacher's checklist of classroom misbehavior and a parent questionnaire used in the National Educational Longitudinal Study of 1988; National Center for Education Statistics, n.d.-a), and teacher and parent rating forms for students found in RAND's Education Assessment Finder (Hamilton et al., 2018) and CASEL's (2020b) SEL Assessment Guide, as well as for large-scale assessments such as the PISA Parent Questionnaire (OECD, 2017b), and the CORE Districts Teacher Questionnaire (West et al., 2018).

It is useful to consider the constructs and measures used. ETS's Personal Potential Index comprised six scales with a text box for each scale, which allowed evaluators to justify or elaborate on the ratings. The six scales were knowledge and creativity, communication skills, teamwork, resilience, planning and organization, and ethics and integrity.

Each scale comprised four items. For planning and organization, the items were sets realistic goals, organizes work and time effectively, meets deadlines, and makes plans and sticks to them. The rating scale included six categories: "below average," "average," "above average," "outstanding (top 5%)," "truly exceptional (top 1%)," and "do not have sufficient information to evaluate."

The Swedish military interview (Lindqvist & Vestman, 2011) was based on a clinical psychologist's analysis of several attributes, combined to form a single composite score. The positive attributes were willingness to assume responsibility, independence, outgoing character, persistence, emotional stability, initiative, and social skills. Also considered was ability to adjust to the specific requirements of life in the armed forces (loss of personal freedom). Specific traits screened out were difficulty accepting authority, violent or aggressive behavior, psychopathology, and antisocial tendencies.

The CORE Districts teacher ratings were only for the self-management and social awareness scales, presumably because self-efficacy and growth mindset are largely unobservable. There were nine self-management questions and eight social awareness questions. The teacher questions were in some cases nearly identical to the self-assessment versions (student self-assessment: "I remembered and followed directions" versus teacher rating: "remembered and followed directions"), but in some cases there were small discrepancies (student self-assessment: "I was polite to adults and peers" versus teacher rating: "got along with others"). The response categories for the two scales were the same and were the same for teachers and students ("almost never," "once in a while," "sometimes," "often," "almost all the time").

Others' ratings have several conceptual advantages over self-ratings. An external other can place an observed target (a student) in a distribution of other observed targets (other students), whereas a self-rating is subject to biases discussed in the previous section (e.g., social desirability). Also, two or more independent raters can rate a target, whereas only one self exists (there can be different imagined selves, as in faking experiments).

Analysis, Scoring, Scaling, and Reliability

Almost all analysis, scoring, and scaling of ratings that appear in the literature is based on simple sum scoring across raters (R1, R2) and classical test theory for evaluating ratings. Often, a statistic to measure rater agreement is computed, such as percent agreement, or Cohen's (1960) kappa, which adjusts for agreement due to chance. This can be generalized to the situation with multiple categories (as with Likert scales), where closeness is credited using the weighted kappa statistic (Cohen, 1968), and with multiple raters (Conger, 1980). Although these methods remain popular, Gwet's (2014) AC1 and AC2 statistics may be useful alternatives in some cases (Keener, 2020).

Generalizability theory (Brennan, 2001; Cronbach et al., 1972; see also Lee & Harris, this volume) can be used for analysis of ratings data. Generalizability theory is an extension of classical test theory in which measurement error is not unitary but is partitioned into sources such as multiple items, occasions, raters, or other factors. A generalizability theory analysis was conducted on the Personal Potential Index by McCaffrey et al.

(2018). They found high correlations among the six scales ($r = .67$ to $r = .90$), consistent with a halo effect (Thorndike, 1920). They found a large rater effect, larger than the effect for persons being rated. This is consistent with research (Baxter et al., 1981) finding more similarity between two ratees rated by the same rater than between two rated by different raters. Consequently, McCaffrey et al. (2018) estimated that 10 raters or possibly more might be needed to obtain a score that had the same level of reliability as commonly used admissions assessments. The authors also point out that there can be response style effects operating on evaluators, as there are on self-raters, including severity and leniency tendencies. These factors (halo, lack of rater agreement, rater response biases) likely operate in all rating situations but are only apparent when ratings are quantified.

IRT approaches can evaluate raters, most notably with the Rasch-based Facets software (Linacre & Wright, 2002), which has similar purposes as a generalizability theory analysis. Facets is designed to analyze data from multiple items and raters and estimates rater leniency-severity effects. To our knowledge, Facets has primarily been applied to ratings of cognitive tasks (e.g., writing), rather than to Likert-style ratings of noncognitive competencies.

Validity

EVIDENCE BASED ON CONTENT The content analysis case for ratings by others is essentially the same as that for self-ratings—ratings are straightforward statements expressing the rater's belief about the target based on observations. Others' ratings better predict outcomes than self-ratings with traits high in evaluativeness, such as intellect (Vazire, 2010). The key challenges to self-ratings, calibration and faking, are generally less applicable to observer ratings. An observer can compare the target to others without self-serving bias, increasing calibration, and the observer typically has minimal incentives to fake (there may be special circumstances where this is not true, such as the professor who takes pride in their students' accomplishments). Leniency severity and halo bias exist. Also, calibration is not guaranteed (McCaffrey et al., 2018). Some not-easily-observable traits are known only by the self (Vazire, 2010).

EVIDENCE BASED ON CORRELATIONS WITH EXTERNAL VARIABLES Several meta-analyses have evaluated others' ratings for predicting school (Connelly & Ones, 2010; Poropat, 2014) and workforce outcomes (Connelley & Ones, 2010; Oh et al., 2011). All have shown that ratings by others have higher correlations with outcomes than do self-ratings. Poropat (2014) examined Big Five factors' prediction of GPA. Targets ranged in age, but were mostly college students. The Big Five was measured by various instruments under various design conditions (one versus multiple raters; teacher vs. peer vs. spouse ratings). The main finding was that the correlations with performance given by others' ratings were higher than those given by self-reports for all of the Big Five factors (other rating $\rho = .38, .28, .18, .10, .05$ versus self-rating $\rho = .22, .09, .00, .06, -.02$ for Conscientiousness, Openness, Emotional Stability, Agreeableness, and Extraversion, respectively; Openness correlations go down substantially

when cognitive ability is controlled, but the others do not change). Poropat (2014) claimed that the magnitude of these correlations puts them among the largest effect sizes in education considering Hattie's (2009) meta-analysis of effect sizes in education. Feng et al. (2022) similarly found that teacher reports were more reliable than children's self-reports and more predictive than self-reports of school outcomes controlling for cognitive ability. These studies are generally low-stakes applications, in which ratings, by teacher, guardian, or self, are for school monitoring or program evaluation. McAbee and Connelly (2016) provided a framework for understanding discrepancies between rating sources.

Connelley and Ones (2010) also included school outcomes in their meta-analysis of mostly workplace outcomes. They found that simple correlations between self- and others' ratings of outcomes were similar, but others' ratings were substantially greater than and incremental to self-ratings in predicting outcomes when unreliability was corrected for. In addition, they conducted a meta-analysis on interrater agreement and found, like McCaffrey et al. (2018), that many raters, up to 10, would be needed to get reasonable interrater reliability; at the very least, they recommended having more than 1 rater.

Oh et al.'s (2011) meta-analysis focused on workplace outcomes. They found predictions comparable to the other meta-analyses. The true-score correlations between predictors and outcomes with other reports were ρ = .37, .26, .21, .31, .27, and for self-reports they were ρ = .22, .05, .14, .10, .09, for Conscientiousness, Openness, Emotional Stability, Agreeableness, and Extraversion, respectively. Other report correlations were like Poropat's (2014) estimates, except for the latter two (Agreeableness and Extraversion), which were higher in Oh et al. (2011) (studies did not overlap because Poropat, 2014, used educational outcomes, whereas Oh et al., 2012, used workforce outcomes). Oh et al. (2012) showed that multiple raters improved prediction (e.g., for Conscientiousness, ρ = .32 and .41 for one and three raters, respectively); however, even correlations with outcomes based on single raters were substantially higher than those based on self-ratings. Oh et al. (2012) also conducted regression analyses to show that others' ratings added incremental R^2 prediction to self-ratings (on average, ΔR^2 = .048) but that the reverse was not true (on average, ΔR^2 = .005). An earlier meta-analysis (Harris & Schaubroeck, 1988) estimated correlations of performance ratings between self–supervisor (ρ = .35), self–peer (ρ = .36), and peer–supervisor (ρ = .62), indicating the self is the outlier.

Two other studies from the economics literature similarly reported high correlations with outcomes based on others' ratings. The Swedish enlistment study (Lindqvist & Vestman, 2011) examined earnings, employment, and chronic unemployment 20 years after an overall noncognitive rating by one clinical psychologist was collected from 18-year-old males (the rating factors are discussed in a previous section). They found that the psychologist's noncognitive rating was a stronger predictor of outcomes than cognitive ability, particularly for the outcome of chronic unemployment, where it was a stronger predictor than cognitive scores by a factor of 5.

The National Educational Longitudinal Study of 1988 teacher ratings study (Segal, 2013) examined earnings 20 years after two eighth-grade teachers' ratings of student misbehavior (dummy variable: 1 if either teacher rated "yes" to any of the items: *rarely completes homework, is frequently absent, is frequently tardy, is consistently inattentive in class, and is frequently disruptive*; 0 otherwise). She found that after controlling for test scores, eighth-grade misbehavior predicted earnings, and after controlling for educational attainment, it predicted earnings at all educational levels, whereas achievement predicted earnings only for males with postsecondary degrees.

A question is, How much do others' ratings correlate with self-ratings? West et al. (2018) presented correlations between teacher and student self-ratings of $r = .41$, .38, and .47 for elementary, middle, and high school students, respectively, for self-management and $r = .22$, .21, and .26 for social awareness (teacher ratings were averaged when there was more than one teacher). However, correlations between teacher ratings on the two scales were considerably higher, $r = .85$, .82, .85, for elementary, middle, and high school teachers, indicating a considerable halo effect! Connelly and Ones (2011) also found that self–other agreement was at least moderately high. The mean observed self–other correlations were $\bar{r} = .37$, .34, .34, .29, and .41, for Conscientiousness, Openness, Emotional Stability, Agreeableness, and Extraversion, respectively, boosted substantially when correcting for unreliability. West et al.'s (2018) self-management and social awareness are similar to Connelly and Ones's (2011) Conscientiousness and Agreeableness, respectively, and then the correspondence in findings is much clearer. Also, Connelly and Ones (2011) found that the most accurate others' ratings came from friends and family (they did not examine teachers as a separate category); but there were some traits, such as Extraversion, for which strangers' ratings were as accurate as ratings from those who had known the target longer.

Fairness

The use of differential item functioning and measurement invariance approaches for analyzing ratings data is extremely rare, partly reflecting the rarity of ratings data for noncognitive characteristics generally. Also, response style effects and social desirability could undoubtedly be found in ratings data, but they have not been systematically investigated. Two other rating biases have been examined. One is the halo effect (Thorndike, 1920), the lack of differentiation in ratings across constructs. A halo effect was observed in the results of West et al. (2018), as the extremely high intercorrelations among ratings on the four CORE Districts variables (all $r > .80$). Similarly, McCaffrey et al. (2018) showed very high correlations among the six Personal Potential Index ratings (median $r = .81$). The other bias is rater leniency versus severity, which is like acquiescence response style in self-ratings. We are not aware of a body of literature that has examined this phenomenon in noncognitive ratings, although it is well studied and documented in ratings of cognitive test performance (Linacre & Wright, 2002).

In summary, others' ratings likely overcome some problems with self-ratings, most notably the reduction in social desirability, faking, response style, and calibration

biases. This may be what is responsible for findings of increased prediction of outcomes compared to self-ratings, which was found in several meta-analyses of both school and workforce outcomes. However, others' ratings introduce some additional problems, particularly halo, which results in a lack of differentiation between constructs, and the low level of agreement between raters. A common recommendation is to include others' ratings where possible, particularly more than one rater, with an awareness of their limitations.

Anchoring Methods

Anchoring methods refer to a set of techniques for increasing rating scale comparability across targets using scale anchors. These methods are intended not to mitigate or adjust for faking, but to increase interpretability of responses, particularly in low-stakes assessments. A scale anchor is a concrete behavioral descriptor or vignette, which may be more resistant to response biases than response category labels like "agree" or "often." Examples of anchoring methods are anchoring vignettes (AVs; King & Wand, 2007), behaviorally anchored rating scales (BARS; Kell et al., 2017; Klieger et al., 2018), behavioral summary scales (BSS; Borman, 1979), Rasch/Guttman scenario scales (Ludlow et al., 2020), and perhaps single-item measures of personality, bipolar single-item scales with richer descriptors (Woods & Hampson, 2005), which could have been included here, but we do not discuss them in detail. In principle, all of these can be used as self or other ratings, although AVs and single-item measures of personality are typically used as self-ratings and BARS, BSS, and Rasch/Guttman scenario scales are typically others' ratings.

Anchoring Vignettes

AVs supplement self-rating scale measures with additional vignettes that describe one or more hypothetical persons—the respondent's task is to rate vignettes the same way they rate themselves (on the same rating scale, with the same stem). With more than one vignette, the vignettes are designed to be at different locations on the scale (e.g., a high and low location). Assuming vignettes are interpreted the same way by all respondents (vignette equivalency assumption) and that the rating standards for self and vignettes are the same (response consistency assumption), self-ratings can be interpreted in relation to the vignette ratings to create a new anchored self-rating scale. Consider a case with one vignette ($J = 1$) describing a hypothetical person, X:

> X sometimes arrives late to class; usually, but not always completes their homework; and sometimes procrastinates on assignments. How much do you agree that X is hardworking? (strongly disagree, disagree, agree, strongly agree).

The respondent is also asked to rate themselves on the same item:

> How much do you agree that you are hardworking? (strongly disagree, disagree, agree, strongly agree).

The two responses enable transforming the Likert self-rating to a new relative (to the vignette) rating called C_i, for each respondent *i*. Given an item and the self- and vignette rating on that item, the respondent might rate themselves higher, lower, or at the same level as they rated the vignette; thus, the new scale C_i is a three-category scale. With two vignettes ($J = 2$), there are more possibilities (rate self higher, the same, or lower than the high-location vignette and higher, the same, or lower than the low-location vignette), and with three vignettes, still more (the number of categories in C is $2J + 1$). Response ties (rating the vignettes the same) and misorderings (rating the lower of two vignettes higher than the higher vignette) reduce the number of transformed categories (e.g., higher than both, lower than both, or neither), but do not entirely invalidate the nonparametric transformation rule logic.

SCORING AVs can be scored in two ways. A nonparametric approach transforms self-ratings to a new relative-to-anchor scale, as described in the previous paragraph. Some self-rating responses cannot be treated in a straightforward way because of ties or misorderings. Vignettes are written to be at a particular location (e.g., high and low), but they can be moved after the fact based on empirical findings as long as they are ordered the same for all respondents. Self-rating responses under these conditions can be treated as missing or by various assignment rules, which can be specified (Wand et al., 2016).

Parametric AV scoring uses ancillary variables such as country, gender, and race/ethnicity in addition to self- and vignette ratings to estimate scores on the latent factor, the unobserved perceived level. The parametric approach has several advantages over the nonparametric approach: Not all respondents have to rate all the vignettes, and misorderings are assumed to result from random error. The R package (R Core Team) "anchors" (Wand et al., 2011) handles both nonparametric and parametric scoring and includes many options and diagnostics.

RELIABILITY Internal consistency reliability estimation with AVs is not straightforward. Von Davier et al. (2018) showed that Cronbach's alpha is not an appropriate measure of internal consistency for AVs, particularly when ties and misorderings occur, and that scale intercorrelations will increase when corrected by the same set of vignettes. However, test–retest correlations would still be useful (with the same or different sets of vignettes), as would a generalizability theory (variance component) analysis using vignettes as a random factor (e.g., using two sets of vignettes for the same construct; on one or multiple occasions). To our knowledge, such research has not yet been reported.

VALIDITY AND FAIRNESS The justification for AVs is that they adjust responses for response style biases and reference group effects with information from responses to the vignettes. This assertion depends on the vignette equivalency (different respondents see the vignettes the same way) and response consistency (respondents apply the same standards in rating themselves and the vignettes) assumptions. These can be vio-

lated (Kapteyn et al., 2011), and methods for evaluating these assumptions are needed. Despite this, a major finding with AVs is that they have been found to address the "attitude–achievement paradox" and increase country comparability on certain noncognitive scales (Kyllonen & Bertling, 2014). The attitude–achievement paradox is the common finding of a positive attitude–achievement correlation within country (e.g., between a personality factor and achievement), with a simultaneous negative correlation at the country level (e.g., the countries with high average personality scores will have low average achievement scores). This is a recurring finding in large-scale international assessments for many (but not all) scales, particularly ones that are more abstract, such as personality scales. Kyllonen and Bertling (2014) and He et al. (2017) showed on PISA 2012 that AV-adjusted scores on two scales (teacher support and classroom management) had higher correlations with mathematics achievement than unadjusted scores. He et al. (2017) also found that Asian cultures topped the ranking on teacher support after adjustment, a finding more in line with the literature. AV approaches have now been tried in numerous studies, with mixed results (Mõttus et al., 2012; Primi et al., 2016). It is a complex method and there are many issues to consider in vignette and study design, but the findings from PISA 2012 seem promising and worth pursuing.

Rasch/Guttman Scenario Scales

Rasch/Guttman Scenario Scales is a method developed by Ludlow and colleagues (Ludlow et al., 2019, 2020), which also use vignettes, called scenarios. The respondent is presented a series of scenarios and asked to rate themselves (on a five-category scale going from strongly higher to strongly lower) relative to the character described in the scenarios; the scenarios describe people at different locations on the construct. Scenarios are developed using Guttman's facet theory and mapping sentence approach (Guttman & Levy, 1982). This differs from AVs in that with AVs, the respondent rates self and vignette independently. Here, one compares oneself directly with the vignette (scenario). Constructs are defined hierarchically, so goal orientation might be a higher order construct, with clarity, effort, and frequency as lower order facets (analogous to the facets–factor relationship in the five-factor model; John & Srivastava, 1999). The Sentence Map contains sentences with location fillers representing different construct levels. Ludlow et al. (2020; Table 1b) provided an example:

> clarity: "<name> has a {not at all clear| somewhat clear| extremely clear} vision of how to make goals a reality"
> effort: "he/she places {almost no effort |some effort |a tremendous amount of effort} toward making long term aims a reality"
> frequency: "he/she {almost never |sometimes |almost all the time} spends time in the day engaged in activities that bring him/her closer to his/her goals"

Because there are three mapping sentences with three locations each, there are m^n (m levels, n sentences) possible sentence combinations (e.g., LLL to HHH), but only enough need to be written (Ludlow et al., 2020, wrote 7 of the 27) to ensure spread

across construct levels. The sentences are transformed to scenarios through editing for flow and naturalness; some facets might be combined in a single sentence. Ludlow et al. (2020) (Study 2) presents six goal orientation scenarios: the highest (HHH) is "Bill knows how to make his goals a reality and constantly exerts tremendous effort toward accomplishing them"; the next highest (HHM) is "Jill is sure she knows how to achieve her goals. She is actively engaged in efforts to make her goals a reality and wishes she could be even more engaged." The second to lowest level scenario (LML) is "Jim is unclear on how to make his plans a reality. Although he is rarely engaged in activities that move him closer to his goals, he does put some effort into working toward them." Pilot testing can result in rewriting or swapping out scenarios (e.g., an LLH for an LHL).

SCORING AND RELIABILITY Ludlow et al. (2019, 2020) used the Rasch rating scale model for scaling and scoring. The analysis provides item and person locations on the variable maps, which can be examined to ensure good coverage on the construct continuum; revisions to items or to the selection of items can be made to ensure good coverage, as Ludlow et al. (2020) demonstrated. They also showed good model fit for their measure.

VALIDITY AND FAIRNESS The justification for using this assessment to draw inferences about an individual's trait level follows the basic logic of other self-ratings measures—an individual is making claims about themselves. The claim is bolstered compared to simple self-ratings in that the self-description is in relation to a hypothetical other, although here, the claim is perhaps more direct than AVs; respondents are not rating themselves, then a vignette; respondents rate themselves directly compared to the vignette. This would seem to make the response consistency assumption more likely to be satisfied. A finding of good model fit supports the justification for its use. Poor model fit would indicate potential inconsistencies in the judgments being made by the respondent. Like AVs, scenario scales should reduce response style biases and reference group effects, although in Ludlow et al.'s (2019, 2020) application the rating scale (strongly higher to strongly lower) could allow response style effects.

The method is relatively recent, and little work has been done thus far on group differences or measurement invariance. However, measurement invariance approaches for Rasch measurement are available (Millsap, 2011; Wu & Estabrook, 2016).

BARS/BSS

BARS (Kell et al., 2017; Klieger et al., 2018; Oswald et al., 2004; Shultz & Zedeck, 2011) are discrete category rating scales (e.g., 0 to 5) with behavioral descriptions based on critical incidents appearing at various points along the rating scale continuum. BSS are similar, but represent composites of several incidents, resulting in more abstract behavioral descriptions. The behavioral descriptions, or anchors, help raters calibrate their ratings. The rater can locate the target relative to the anchors by essentially doing mental paired comparisons (e.g., "*x* is higher than Anchor 1,

but not as high as Anchor 2). Computer-adaptive versions of BARS, called CARS, have also been developed and evaluated with some success (Darr et al., 2017). A related alternative, the relative percentile method (Goffin et al., 2009), asks raters to rate employees against other employees, so that employees themselves serve as scale anchors.

BARS are developed through the initial collection of critical incidents, settings in which the construct is expressed through behavior. For example, Klieger et al. (2018) asked subject matter experts to recall examples of "highly ineffective, just good enough, and highly effective behavior" on job domains such as initiative/work ethic and flexibility/resilience. Subject matter experts are asked to describe the situation, the behavior, and the outcome associated with the critical incident and sometimes an interpretation (Hall et al., 1995). Incidents are sorted by dimensions or themes and then edited into behavioral statements such as "prior to transitioning to a new department, reaches out to relevant coworkers to inquire about strategies for a new position" (Klieger et al., 2018). Incidents are then rated on a six-point effectiveness scale resulting in a BARS. This general method was applied to lawyering (24 job domains; Shultz & Zedeck, 2011).

Scoring and Reliability

The fielded BARS provide ordered categorical data, typically treated in classical fashion, using sum scores and classical reliability theory. The method naturally lends itself to a generalizability theory analysis, with raters, items, settings, and occasions potentially serving as design facets over which to evaluate generalizability (Medvedev et al., 2019; Ohland et al., 2012). To our knowledge, except for Darr et al. (2017), few efforts have developed IRT models for the BARS development or application cycle.

Validity and Fairness

As an outcome measure, BARS do not have the correlational evidence that typically comes with predictor measures. Instead, for BARS, like licensure tests, validity evidence tends to be associated with the process by which the assessment is developed and based on content matches between the assessment items and the intended constructs (i.e., evidence based on test content). The BARS development process (Kell et al., 2017; Klieger et al., 2018)—the identification of the construct(s) based on job analyses, the collection of critical incidents, effectiveness evaluations, and content evaluations of items—is rigorous and designed to ensure validity evidence. Debnath et al. (2015) argued that BARS remain popular appraisal instruments because of their advantages over instruments that focus on traits (because traits are relatively immutable and subject to halo effects), results (because hard criteria are unavailable for many jobs), and behavior (such as checklists; they are too rigid and leave out evaluator judgment).

Neither differential item functioning nor measurement invariance analyses tends to be performed on BARS. However, self–other similarity bias exists for ratings. Supervisor–subordinate racial differences are associated with lower ratings for both Black and

White subordinates (Elvira & Town, 2002) and sometimes gender similarity effects are found, although not universally (Stone et al., 2016).

Forced-Choice and Ranking Methods

Forced-choice and ranking methods refer to tasks in which respondents are asked to choose from, rank, or express preferences between two or more statements. Ranking is an alternative response format to ratings. A forced-choice example would be, "Choose the statement more like you: 'I enjoy working with others,' or 'I am a hard worker.'" A ranking example would be, "Rank the following 1 to 4 for how they describe you (1 = *most like you*; 4 = *least like you*): 'I enjoy working with others,' 'I am a hard worker,' 'I am calm even in stressful situations,' 'I enjoy speculating about the origins of the universe.'"

Design Issues

UNIDIMENSIONAL VERSUS MULTIDIMENSIONAL FORCED CHOICE Forced-choice items may be unidimensional or multidimensional. Forced-choice statements on a unidimensional item are drawn from the same dimension but vary in their location: "I am a hard worker" versus "I do enough to get by." The Myers–Briggs Type Indicator (Form M) uses 93 *unidimensional* forced-choice items (statement pairs), each asking individuals whether they are better described by positive or negative pole descriptors for the four bipolar factors they measure (extravert–introvert, sensing–intuition, thinking–feeling, judging–perceiving; roughly 23 pairs per factor). SHL's Operational Personality Questionnaire (OPQ32r; SHL Group, 1999; Brown & Bartram, 2009) uses *multidimensional* forced-choice items (three-statement triads) to measure 32 dimensions, with items such as "Please choose one MOST true and one LEAST true statement: 'I enjoy the companionship of others,' 'I try out new activities,' 'I look to the future.'"

BLOCK SIZE Block size refers to the number of statements ranked, typically either two, three, or four, with exceptions (Dueber et al., 2019, used 12). SHL's two OPQ versions differ by having either three (104 triads; OPQ32r) or four (104 tetrads) statements (OPQ32i) per block (Brown & Bartram, 2013; the OPQ32r's triads are the same as the OPQ32i tetrads with one statement removed per block). ETS's FACETS (12 to 15 dimensions, 104–120 pairs; Naemi et al., 2014), DoD's TAPAS (22 dimensions; Drasgow et al., 2012), and the Myers–Briggs Type Indicator Form M all use pairs.

RESPONSE FORMAT Common response formats are PICK (choose the option "most like me"), MOLE (choose the option most like me and the one least like me), and RANK (rank options 1 to *n*) (Hontangas et al., 2015). Some formats allow a graded preference, such as strongly prefer *i*, slightly prefer *i*, prefer neither *i* nor *k*, slightly prefer *k*, or strongly prefer *k* (Gallup's Clifton StrengthsFinder 2.0, Asplund et al., 2014; 177 pairs to measure 34 dimensions), which can be modeled using IRT (Brown & Maydeu-Olivares, 2018). Another format is compositional (Aicheson, 1982; Brown,

2016), in which respondents express choices by dividing (say) 100 points across a set of four descriptors, such as assigning 50 to dependable, 20 to curious, 20 to modest, and 10 to calm; or selecting 10 descriptors from a pool of 30.

RESPONSE PROCESS (DOMINANCE VERSUS UNFOLDING) For a dominance response process, the higher a respondent is on a trait, the more likely they will endorse the trait (e.g., respond "strongly agree" or "almost all the time"). This is the typical assumption in personality assessments; it enables data to be fit by CTT models and the same IRT models used in cognitive testing, where the higher the ability, the more likely one is to get the item correct. An unfolding response process involves an ideal point in which a respondent is most likely to endorse an item whose location is the same as the respondent's latent trait position and increasingly less likely to endorse an item farther away from their trait location, allowing rejection "from above" or "from below" (Chernyshenko, 2003; J. S. Roberts et al., 2000). An introvert would reject the item "I enjoy talking to a friend in a quiet cafeteria," preferring no social interaction, and an extravert might reject the same item, preferring a more intense social experience. Not common in personality, ideal point models are common in attitude measurement, allowing disagreement with a middle-of-the-road policy from the ideological left or right. The appropriateness of unfolding for personality measurement is open to debate (Drasgow et al., 2010; Oswald & Schell, 2010). Nevertheless, using simulated and real data, Fu et al. (2022) found support for the theoretical claim but found mixed evidence with real data. They recommended fitting different models to different statements.

DATA TYPE Data yielded by forced-choice assessments can be either ipsative, quasi-ipsative, or normative (Salgado et al., 2015). Ipsative data produced by a forced-choice measure refers to a measurement scale within a single person, not comparable to other people (Cattell, 1944). Clemans (1966), Hicks (1970), and Gleser (1972) documented the qualities and limitations of this kind of data type, such as constraints on the average correlation among scales ($\bar{r}$ is bounded by $-1/(d-1)$ and $(d-4)/d$, where d is the number of dimensions) and average correlation of scales with outcome ($\bar{r}_{xy} = 0$), and its consequent unsuitability for factor analysis and internal consistency reliability computation. Normative data refer to a measurement comparable across people, as produced by a Likert scale (or right/wrong data from a cognitive test). Such data are amenable to factor analysis and internal consistency reliability measurement. Quasi-ipsative data refer to data produced by a forced-choice procedure, but with properties not as limited as ipsative data and closer to normative data-like qualities. In the literature there are several approaches for generating quasi-ipsative data. One is using forced-choice methods scoring only a subset of the dimensions. As d increases, the limitations of ipsative data are less restrictive. A variant is the use of forced-choice methods with phantom dimensions, where the phantom (unscored) dimension is represented by a set of items that are not related to the targeted (scored) dimensions (Horn, 1971; Salgado & Lado, 2018). Another approach is the IRT methods developed to overcome the limitations of ipsative data (Brown & Maydeu-Olivares, 2013; Stark et al., 2005).

Modeling, Scoring, and Scaling

UNIDIMENSIONAL APPROACH Unidimensional pairwise preference task data are normative data and can be analyzed by classical test theory and sum score approaches and by IRT models for binary and polytomous data (Rasch [von Davier, 2016], Rasch rating scale [Andrich, 2016], Rasch partial credit [Masters, 2016], two-parameter logistic, 2PL [Birnbaum, 1968], generalized partial credit [Muraki & Muraki, 2016], and the graded response models [Samejima, 2016]). Multidimensional items with unscored dimensions (phantom dimensions) can also be analyzed using unidimensional methods if targeted dimension statements are only paired with unscored dimension statements (Horn, 1971; Salgado & Lado, 2018). Such assessments yield binary or polytomous data that reflect preference (or degree of preference) for the target dimension only (i.e., not also for the unscored dimension). These methods are easy to implement, although not as efficient, perhaps, as multidimensional methods.

UNFOLDING APPROACH The unfolding approach proposed by Stark et al. (2005) involves two steps, a pretesting calibration step with Likert rating responses, followed by the test administration with pairs. The data from the Likert rating pretesting are calibrated based on the generalized graded unfolding model (GGUM; J. S. Roberts et al., 2000). Pairs are assembled so that the two statements are from different dimensions (mostly) and are matched on social desirability. Social desirability can be based simply on the Likert statement mean or principal component score after negative keyed items are reflected (Bäckströmm & Björklund, 2013; Kuncel & Tellegen, 2009) or from a separate data collection with "fake good" instructions (Pavlov et al., 2021). Test administration provides binary data on the pairs (i.e., selecting statement i over k or not). The binary choice data is analyzed with the multi-unidimensional pairwise preference (MUPP) model. Stark et al. (2005) argued that separating the Likert rating calibration step from the analysis of pairs has advantages in simplifying computation and in having a calibrated statement pool that can be used to create adaptive tests.

However, the Stark et al. (2005) method relies on the assumption that the GGUM calibrations are invariant across the Likert and forced-choice format, even though the two administrations would typically involve different samples in different contexts perhaps even years apart; and the Likert statement stands by itself, whereas the forced-choice statement has a context of other statements in the item block. Also, Stark et al.'s approach does not lend itself to item analysis based on the blocks. So, P. Lee et al. (2019) proposed GGUM-RANK to estimate statement parameters and person parameters directly from the test administration of multidimensional forced-choice blocks without pretesting. It uses a generalization of the GGUM model proposed by Hontangas et al. (2015) in which ordering (e.g., A > B > C) is seen as a sequence of steps, selecting A versus B and C, then selecting B versus C. This enables the use of

triads rather than pairs (as in Stark et al., 2005). Hontangas et al. (2015) also showed that ranking methods were superior to other approaches and demonstrated how true scores, expected a posteriori (EAP) estimates, and traditional scores could be obtained and compared. P. Lee et al. (2019) demonstrated their approach in both simulations and with empirical data using Markov chain Monte Carlo (MCMC) estimation with a Metropolis–Hastings within Gibbs (Tierney) algorithm. They seem to have found that the triad format was more efficient than pairs, that statement discrimination within blocks was critical for estimation accuracy, and, most important, that the method performed well in providing convergent and discriminant validity evidence with external measures.

Thurstonian IRT and Other Dominance Approaches

Brown and Maydeu-Olivares (2011) and Maydeu-Olivares and Brown (2010) proposed Thurstonian IRT as a dominance alternative to the unfolding approach, basing it on Thurstone's (1927) law of comparative judgment. The method does not require pretesting and can be applied to forced-choice or ranking data of any size blocks. Response data are converted to $n(n-1)/2$ binary pairwise preferences for the n statements in a block and analyzed. Brown and Maydeu-Olivares (2012) provided a tutorial on how to estimate the model using a confirmatory factor analysis approach with Mplus (Muthén & Muthén, 2001).

The basic concept is that each statement i elicits a latent utility, t_i (i.e., a degree of *desiredness* or *endorsement-worthiness* to respondents), and statement i is preferred over statement j if $t_i \geq t_j$. Utilities are modeled as $t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i$, where μ_i is the latent utility mean, λ_i is the factor loading of item i on the latent trait η_a, and ε_i is a normally distributed error with mean 0 and variance ψ_i^2. The latent response variable (i.e., difference in utilities) is modeled as $y_{ij}^* = t_i - t_j = \mu_i - \mu_j + \lambda_i \eta_a - \lambda_j \eta_b + \varepsilon_i - \varepsilon_j$, which involves two latent traits (η_a, η_b), location parameters $(\mu_i - \mu_j)$, factor loadings $(\lambda_i - \lambda_j)$, and uniqueness parameter $(\psi_i^2 + \psi_j^2)$. All these are specified as constraints in fitting a confirmatory factor analysis model in Mplus. Because of the complex error structure and often large number of latent traits, Thurstonian IRT models are typically estimated using limited-information methods.

Variations have been proposed. Wang et al. (2017) argued for the advantages of a one-parameter Rasch ipsative model (the Thurstonian IRT model is a two-parameter $[\mu, \lambda]$ model): It yields a single utility value for each statement, allows for between- and within-individual differentiation, and does not rely on statement blocks requiring mixed location (positive, negative) statements as Thurstonian IRT models do. This is an important issue because it can be a limitation of Thurstonian IRT models for high-stakes use (Bürkner et al., 2019). Morillo et al.'s (2016; Hontangas et al., 2016) MUPP-2PL model uses the MUPP choice model, but the 2PL instead of GGUM for calibration, making it a dominance model. Bunji and Okada's (2019) D-diffusion Thurstonian IRT model incorporates

response time information in the forced-choice item responses to improve model fit and increase the outcomes predictions. Response time information can inform response models (van der Linden, 2007); in personality, the assumption is that fast selection indicates easier option choices, such as stating "agree" to a binary Likert question (Ranger, 2013) or selecting an option in a forced-choice application (Bunji & Okada, 2019).

Reliability

Estimating reliability for unidimensional forced-choice models is straightforward, whether using CTT (e.g., alpha) or IRT (marginal reliability) approaches. Estimating reliability for the Stark et al. (2005) model is not straightforward. Hontangas et al. (2015) used a simulation study with known true scores and correlated them with EAP and traditional scores to show that EAP scores were more reliable. However, they did not address how reliability could be computed without knowing true scores. Seybert and Becker (2019) computed test–retest reliabilities, finding them to be only moderate and slightly lower than Likert scale versions of items. However, with rating scale test–retest reliability approaches, what is being measured is a mix of trait stability and stability of construct-irrelevant factors such as response style. Forced-choice methods are intended to reduce the proportion of response style variance and therefore should have lower test–retest factor correlations.

Estimating reliability with multidimensional forced-choice measures is discussed by Brown and Croudace (2015). They introduce the marginal reliability concept which can be estimated as theoretical or empirical reliability. Theoretical reliability involves averaging the model-based squared standard error for all trait values in the theoretical distribution, which requires multidimensional integration. Empirical reliability is preferred with many dimensions. This involves averaging squared standard errors of estimated trait scores in the sample, depending on the estimator used. For EAP, this is the posterior likelihood variance; for Maximum Likelihood (ML), the Fisher information inverse; for maximum a posteriori (MAP), the posterior information inverse. Brown and Maydeu-Olivares (2018) provided the solution for computing reliability from matrices produced by Mplus and provide R code for computing item information, standard errors, and reliability in the online supplement to that article. Mplus beginning in version 8.1 provides standard errors for MAP scores for IRT models of any dimensionality, which can be used to compute empirical reliability.

Validity and Fairness

The purpose of forced-choice and ranking approaches is to minimize the effects of response distortion but without altering the construct score. There are two situations in which this is important: high-stakes assessment, in which socially desirable responding or faking occurs; and for comparing individuals or groups with different response styles.

Faking effects can be examined experimentally: A treated group is asked to "fake good" to increase prospects for getting the job. Score inflation, the difference between scores

for the control and fake good groups, is less with forced choice compared to Likert-style measures (Cao, 2016; Cao & Drasgow, 2019), particularly when statements within a block are balanced in social desirability and normative approaches (e.g., IRT as discussed here) used for scoring. Bartram's (2007) meta-analysis showed that correlations with external measures (manager competency ratings) were higher with a forced-choice (r = .38) compared to a rating scale format (r = .25), in a study where managers used both. (With N = 1,460 and r = .50 across formats, this is a significant difference.)

A meta-analysis by Salgado and Tauriz (2014) showed that the multidimensional forced-choice format had substantially higher correlations with job performance and academic outcomes than the unidimensional forced-choice format. For Conscientiousness, rho (correlation between the predictor and outcomes corrected for predictor and criterion reliability and indirect range restriction in the predictor) was .40 for quasi-ipsative measurement (which includes both IRT and non-IRT multidimensional forced-choice assessments) but only .16 for normative forced choice (i.e., unidimensional). And this value, .40, is larger than any other meta-analyses of these variables, which tend to be dominated by Likert-scale measures.

A caveat to these positive findings is that cognitive ability may moderate these relationships such that brighter individuals may be more able to fake (Christiansen et al., 2005). However, Kyllonen et al. (2020) demonstrated incremental prediction beyond the GMAT/GRE and undergraduate GPA in predicting graduate business school grades and leadership activity for a forced-choice assessment administered during the admissions process. The increment in adjusted R-square was approximately .04 to .07 (depending on analysis details).

Forced-choice methods have advantages for evaluating culture/language group differences. Bartram (2013) found that at the individual level forced-choice and rating scale measures agreed moderately, but at the country level, correlations with external economic variables (the United Nation's Human Development Index, the Global Competitive Index) were more sensible with forced-choice measurement: Forced-choice Conscientiousness showed positive correlations with the economic variables, while rating scale Conscientiousness showed negative correlations. Country-level correlations do not have to follow individual-level correlations (ecological fallacy), but the forced-choice finding is in line with expectation and the rating scale finding is puzzling. In PISA 2012, two forced-choice scales (subject matter preference and learning style preference) were administered in a rating scale format in 2003 (same statement content). The rating scale format led to high correlations between dimensions and near-zero correlations with achievement. The forced-choice format led to expected negative correlations with each other (as an ipsative measure), but math preference had a high correlation with math achievement, as one would expect.

The studies cited here support the conclusion that forced choice improves the quality of information elicited with a survey instrument, at least when steps are taken to address the problems created by ipsative measurement. Designing and scoring forced-choice assessments is challenging—these are complex models relying on many assumptions.

And there are many questions remaining on how to improve the methodology. There are also user experience challenges. Respondents often complain about having to choose between two equally attractive (or unattractive) choices, even if test-taker motivation is unaffected (Sass et al., 2020). Still, forced-choice measurement is a promising methodology that ought to be included routinely in survey assessments of noncognitive skills, either as a supplement to or as a replacement for rating scale measures, particularly for high-stakes and cross-cultural assessments.

Situational Judgment Tests

Situational judgment tests (SJTs; also practical intelligence measures; McDaniel & Whetzel, 2005) have been popular assessment methods for organizational recruitment and selection (U.S. Office of Personnel Management, 2018a). This may be because of their high face validity (compared to standardized tests), low administrative costs (compared to interviews), relatively moderate subgroup differences (compared to standardized tests; Oostrom et al., 2015), lower susceptibility to faking (compared to personality tests; Kasten et al., 2018), and incremental outcome prediction beyond cognitive ability and personality (Clevenger et al., 2001; McDaniel et al., 2007). SJTs are also being considered in education for undergraduate (Oswald et al., 2004), business school (Hedlund et al., 2006), medical school (Patterson et al., 2016), and dental school admissions (Buyse & Lievens, 2011). SJTs can in principle be used to measure various skills but are mostly designed to measure interpersonal skills (Christian et al., 2010). The American Association of Medical Colleges is pilot testing SJTs for medical school admissions to measure eight primarily interpersonal competencies including service orientation, social skills, cultural competence, teamwork, and ethical responsibility (Ellison et al., 2024). SJTs are often ideally suited for education, training, and development purposes as well. Cox et al. (2017) presented evidence for advantages of SJTs over traditional training.

There are many kinds of SJTs, varying by intended construct (Chan & Schmitt, 2005; Christian et al., 2010; McDaniel & Whetzel, 2005), length (Crook et al., 2011), textual versus video items (Chan & Schmitt, 1997), instructions (Ployhart & Ehrhart, 2003), and other features (see reviews by Campion et al., 2014; Lievens et al., 2008; Oostrom et al., 2015; and Whetzel & McDaniel, 2009). Here, we focus on the most important considerations.

Ratings Versus Rankings

SJT items comprise a stem (situation description) followed by a set of response options, typically four to eight, but sometimes as few as one (Crook et al., 2011). Respondents can either rate or rank options. The number of rating categories typically ranges from two (e.g., "effective" vs. "ineffective") to seven or more (e.g., "not at all effective" to "very effective"). Rankings can be full ("rank all options from best to worst"), partial ("choose the best option and choose the worst option"), or single choice ("choose the best option"). Arthur et al. (2014) concluded that ratings were superior to rankings for SJTs based on their higher correlations with personality,

but personality measurement is ratings based and therefore susceptible to spurious associations with ratings-based SJTs. The disadvantages of ratings are discussed in "Self-Report Ratings" and the advantages of rankings over ratings are discussed in "Forced-Choice and Ranking Methods."

Best-Choice (Knowledge, "Should Do") Versus Self-Prediction (Behavioral Tendency, "Would Do") Instructions

Best-choice instructions ask respondents to choose, rank, or rate each option for its effectiveness. Self-prediction instructions ask respondents to choose, rank, or rate options by how they match to what they would do. Best-choice instructions lead to comparatively higher correlations with cognitive ability. Self-prediction instructions lead to comparatively higher correlations with personality (McDaniel et al., 2007). Best-choice instructions are less susceptible to faking than self-prediction instructions (Broadfoot, 2006; Nguyen et al., 2005; Peeters & Lievens, 2005), although responses to what others would do may be even less susceptible to faking (Lievens et al., 2009; Oostrom et al., 2017). Because of their resistance to faking and because knowledge instructions seem to be conceptually more straightforward than self-prediction instructions for eliciting responses, for some applications best-choice instructions may be better suited for SJTs than behavioral tendency instructions from a validity interpretative argument perspective.

Dimensionality and the Constructs Measured by SJTs

SJTs are typically found to be empirically unidimensional, even when constructed to reflect multiple dimensions. Oswald et al. (2004) wrote SJT items reflecting 12 dimensions of college student performance (e.g., leadership, artistic, career, perseverance, ethics), but found a strong general factor accounting for three times the variance of a second factor and only uninterpretable lower order factors. This is common for many constructs in which tests are developed to a framework that specifies several dimensions but a strong general factor accounts for most of the variance among items (Haberman et al., 2009). PISA 2012 specified a number of mathematical content (change and relationships, space and shape, quantity, uncertainty and data) and process (formulating situations, employing concepts, interpreting outcomes) dimensions (OECD, 2013, pp. 28–36), but found little empirical support for their differentiation, instead finding very high latent correlations among process types (all $r \geq .96$) and among content areas (all $r \geq .84$) (OECD, 2014, Table 12.9, p. 231). In such cases, dimensions serve to ensure construct coverage even though empirically they do not identify separate factors. The persistent empirical finding of unidimensionality has prompted discussions of broad factors underlying SJTs. Christian et al. (2010) reported that SJT developers targeted leadership (38% of SJTs), interpersonal skills (13%), personality (10%), teamwork (4%), job knowledge (3%), and heterogeneous composites (33%). However, SJTs might merely reflect broad tacit knowledge and practical intelligence (Chan & Schmitt, 2005; Sternberg & Horvath, 1999) or general domain knowledge (Corstjens

et al., 2017). Responses may alternatively reflect implicit trait policies, beliefs about the effectiveness of responses to situations based on effective trait expression (choosing an agreeable response when the situation calls for Agreeableness) (Motowidlo & Beier, 2010). There have been attempts to address multidimensionality in SJTs through construct explorations (Guenole et al., 2017; Westring et al., 2009). Still, it seems empirically that unidimensionality is found in SJTs even when multidimensionality is intended.

Scoring

Scoring SJTs can be more complex than scoring typical standardized tests: There can be disagreement on the best response to a situation, providing no single correct answer, and response options may vary continuously in appropriateness. SJT scoring approaches have been reviewed focusing on either ranking (Bergman et al., 2006) or rating methods (Weng et al., 2018; Whelpley, 2014). Bergman et al. (2006) discussed empirical keying (based on an outcome measure), theoretical scoring (based on keying options to a theoretical framework), expert scoring (keying to expert judgments), factor scoring, and subgrouping (based on clustering response patterns). Ratings-based scoring methods tend to focus on determining a profile similarity match between a test taker's rating and aggregate expert profile (McDaniel et al., 2011). De Leng et al. (2017) compared 28 ratings-based (profile similarity) measures, varying the reference group (experts versus the test-taking sample), agreement (simple percentage versus consensus), consensus standardization approach (raw/no standardization versus within-person standardization versus rating scale dichotomization), distance metric (absolute versus squared profile distances), and reference group aggregation method (mean, median, mode). Raw consensus, which credits responses by the proportion of the reference group selecting that response (e.g., if 20% of the reference group selects "agree," then respondents who select "agree" are credited .2 score points), led to higher alphas, particularly when the sample was the reference group and their judgments were aggregated using the mean. But Weng et al. (2018) found that raw consensus scoring led to lower correlations with external measures and suggested that extreme responses were problematic and that the value of midrange items was low.

Problems in SJT scoring may partially be attributable to the inherent problems with rating scales. Forced-choice or ranking approaches mitigate these problems and may therefore be better suited for SJTs. Zu and Kyllonen (2020) compared several observed score (including consensus) and IRT scoring methods for scoring forced-choice/ranking SJT methods. They found that consensus methods (using either experts or the sample as the reference group) were the best observed score approaches, but the best overall approach was the NRM with the sample as reference group. Guo et al. (2019) showed that although the NRM is a data-driven method, it produced an invariant scoring key across cohorts. Guo et al. (2016) demonstrated the value of a nonparametric item analysis approach for understanding SJTs, because SJTs often do not have a clear right or wrong answer, and of NRM for scoring.

Reliability

A well-known limitation to SJTs is their low internal consistency reliability (test–retest reliability is higher; Neubauer & Hofer, 2022). Kasten and Freund's (2016) meta-analytic reliability generalization study of SJTs in 271 studies found an average reliability of $\alpha = .61$ (standard deviation, $s = .20$). Factors depressing reliability were high-stakes use, item complexity, high fidelity, fewer items, pick-only response format, and earlier publication dates (i.e., the community is getting better at writing SJTs), but there were no effects for instruction type (knowledge versus behavioral tendency). These findings support the idea that SJTs are typically too short for operational use and that strategies for extracting more information (e.g., pick best and worst) are useful.

Rating measures tend to be more reliable than ranking methods. But reliability of rating methods confounds construct-irrelevant response style variance with construct-relevant trait variance, whereas ranking methods do not. Therefore, this apparent rating advantage cannot be interpreted as providing better measurement.

Validity

There is a strong argument for the validity evidence based on content supporting inferences drawn from SJTs. The development methodology typically involves a careful analysis of the performance situation SJTs are intended to reflect or predict, using critical incidents (see the section on behaviorally anchored rating scales, which uses the same analysis) or related methods to define the intended construct (Sternberg et al., 2000). The method itself asks respondents either what they would do or what the best response to the situation would be; either question should elicit useful respondent information allowing for a straightforward interpretation. The option ratings version of SJTs has the same conceptual problems as rating scales have generally (see the section on rating scales), and therefore the option ranking approaches have interpretive advantages from a validity standpoint.

There are several challenges to the interpretation of SJT responses in a straightforward manner. One is that SJTs ask respondents what they would or should do, but both of those are distinct from an actual response in a real-world situation. Ployhart and Ehrhart (2003) found relatively low correlations between responses in would-do and should-do formats. However, at least with best-choice (knowledge) instructions it could be argued that knowledge is a prerequisite to behavior. With behavioral tendency (would-do) instructions the matter is more muddled because some participants might respond as if to knowledge instructions (should-do) for fear of admitting to poor judgment.

Another challenge relates to a long-standing debate in the literature on what SJTs measure. Oswald et al. (2004) developed SJT items based on 12 college performance dimensions (e.g., leadership, artistic). Interrater agreement on item dimension assignment supported the distinctiveness of 12 dimensions (items with less than 75% agreement were removed), but factor analysis of response data did not, resulting in the adoption of a single score from the measure (Schmitt et al., 2009). Evidence for a general SJT factor is a common occurrence. D. J. R. Jackson et al. (2016) conducted a generalizability theory analysis on three different SJTs and over 10,000 job candidates

finding candidate (person) main effects, with little variation due to dimensions and situations, suggesting a general judgment factor across job situations. For many SJTs the situation descriptions themselves may not be central to what construct is being measured (Schaepers et al., 2020).

A general judgment factor is not the same as the general factor in cognitive tests. SJTs typically add to the prediction given by cognitive tests, as well as to the prediction given by personality measures. McDaniel et al.'s (2001) meta-analysis found a correlation of $\rho = .34$ between SJTs and job performance and $\rho = .46$ between SJTs and cognitive ability; SJTs add incrementally over cognitive ability in predicting outcomes (O'Connell et al., 2007). In education, several studies (Hedlund et al., 2006; Oswald et al., 2004; Schmitt et al., 2009) found that the general judgment factor added to other variables in predicting college outcomes, indicating SJTs are capturing something other cognitive variables are not.

Fairness

A consistent SJT finding is smaller subgroup differences compared to cognitive tests (Arthur et al., 2014; Clevenger et al., 2001; Oostrom et al., 2015; Oswald et al., 2004). There have been few measurement invariance studies of situational judgment, but Prasad (2017) found evidence that interests and socioeconomic status might be related to differential item functioning. For SJTs especially, one might expect culture differences. However, Prasad et al. (2017) found evidence for differences between Chinese versus U.S. respondents on some scales from a college performance SJT, but no differences overall.

Performance Measures of Noncognitive Skills

The assessments reviewed thus far can be called *descriptive assessments*, ones based on how respondents describe themselves (or how others describe them) or what they say they would or should do in a situation. Descriptive assessment approaches are different from cognitive tests in which respondents are asked to perform—to solve problems, recall information, or make a decision—which can be called *performance tests*. There are descriptive cognitive assessments. Mathematics self-efficacy items ("How confident do you feel about having to do the following mathematics tasks? 'solving an equation like $3x + 5 = 17$.' Very confident, Confident, Not very confident, Not at all confident") and mathematics self-concept items ("I learn mathematics quickly. Strongly disagree, disagree, agree, strongly agree"; PISA 2012) are descriptive cognitive assessments. Data from performance tests compared to descriptive assessments undoubtedly inspire more confidence from stakeholders in assertions about student proficiency. This difference between descriptive assessments and performance tests might even account for the slow acceptance of noncognitive assessments despite demand for noncognitive skills. An important question is whether it is possible to make a performance-based noncognitive test.

Performance measures of noncognitive attributes have a long tradition; Cattell and Warburton (1967) referred to them as T-data (test data) measures. French (1948) presented students with a difficult number-series test with instructions stating that some problems had no solutions and that they could receive full credit by marking them as such. It was a disguised persistence test. French found that persistence scores were unrelated to other cognitive tests but were correlated with grades. A modern version is Alan et al.'s (2019) grit game. It comprises two versions of a number grid task in which the goal is to find three pairs of numbers that add up to 100. An easy version has 10 numbers, 5 of which are multiples of 10 and 2 of which are single digit (almost all students successfully completed this task). The difficult version has 20 double-digit numbers and only 1 is a multiple of 10 or 5 (the success rate was much lower). The reward for solving the problem within a short time limit is one gift for the easy version and four gifts for the hard version. Following a treatment designed to improve noncognitive skills by changing students' beliefs about the importance of effort, treated students were found to be more likely to choose the difficult version and 2 years later experienced a $d = .2$ effect on a standardized mathematics achievement test.

Segal's (2012) Coding Speed test, a word-number-lookup-matching task, was administered under low-stakes conditions to 12- to 16-year-old National Longitudinal Study of Youth participants. Segal interpreted the score from this test as a measure of intrinsic motivation because (a) an experiment found that performance improved for a subset of participants when given a financial ($10) incentive, whereas others (the intrinsically motivated) performed at the same level regardless; and (b) test takers taking the test under high-stakes condition outperformed a more highly educated group taking it under low-stakes conditions. Segal found that performance on this test in the low-stakes National Longitudinal Study of Youth condition, controlling for cognitive ability, predicted earnings 23 years later, such that a 1 standard deviation increase in Coding Speed test performance was associated with a 9.5% increase in earnings.

In behavioral economics, *real-effort* tasks, such as Alan et al.'s (2019), are ones on which the respondent works, and the outcome depends on that work. This contrasts with *stated-effort* tasks in which respondents indicate the amount of effort or actions they would take in a situation (Charness et al., 2018). This distinction is analogous to the revealed versus stated preference distinction (Schläpfer & Fischhoff, 2010) and to the distinction we draw here between descriptive assessments and performance tests.

Real-effort task examples include solving mazes and anagrams, adding two-digit numbers, counting zeros in a grid, transcribing meaningless symbols, cracking walnuts, data entry, classifying Amazon reviews, packing quarters into boxes, playing a labyrinth game, digit span, filling envelopes, dragging a ball on a screen, typing alternating keys, and repeatedly typing the same paragraph, among others (Charness et al., 2018). Some of these also measure other factors (cognitive ability, physical skills). However, many could be used as performance tests of factors such as persistence.

Kyllonen and Kell (2018) reviewed tasks that can be understood as performance measures of noncognitive skills. They suggested several categories: cognitive test performance under low-stakes conditions (Segal's 2012 Coding Speed); objective personality tests in the Cattell–Warburton tradition (Kubinger, 2009; Ortner & Proyer, 2015; Ortner & Schmitt, 2014); economic preference tasks, such as time, risk, and social preferences (A. Falk et al., 2015, 2016); and confidence judgments for their calibration to actual performance (Stankov et al., 2013). They also reviewed measures based on test and survey behavior, such as survey effort as indicated by skipping or careless answering on background questions (Hitt et al., 2016; Zamarro et al., 2018), item position effects reflecting less effort on later items (Debeer et al., 2014; Weirich et al., 2017), and hasty response time reflecting lack of effort (Y.-H. Lee & Jia, 2014; Wise & Gao, 2014).

Several other lines of research evaluate performance measures of noncognitive skills. Collaborative problem solving (Hao et al., 2017) combines cognitive and interpersonal skills (OECD, 2017b). A current challenge is assigning credit to individual members (Hao et al., 2019; Martin-Raugh et al., 2020). Idea generation (Bennett & Rock, 1995) can be considered a standard cognitive task, but there are additional dispositional aspects in choosing to respond with more than one solution or more than one answer to a question. Critical thinking also involves dispositional aspects.

Two additional performance measures of noncognitive skills, attitudes, or worldviews have been researched. Implicit association tests (Greenwald et al., 1998) involve interpreting patterns of response times on classification tasks as indicators of underlying attitudes and beliefs. Although mainly known as an implicit bias measure, it can be used to measure personality (Grumm & von Collani, 2007). Conditional reasoning (James, 1998; LeBreton et al., 2020), measures a personality disposition or worldview through multiple-choice answer selection to a reading passage with two correct answers. One answer is consonant with one belief system or worldview and a different answer with a different belief system.

Reliability, Validity, and Fairness

Performance measures are diverse, making summary statements about psychometric issues difficult. However, tasks discussed can be treated with standard psychometric procedures (e.g., IRT; some have, Debeer et al., 2014). Also, fairness concerns are typically ignored. As the use of performance measures increases, we are likely to see increased attention paid to these issues, following guidelines such as those captured in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Challenges associated with these measures include alternative interpretations. Many real-effort tasks likely reflect both the target construct, a personality factor, and construct-irrelevant cognitive or physical ability factors. Watts et al.'s (2018) review of the marshmallow test, an indicator of a child's ability to delay gratification, found that much of its association with later education and work achievements was attributable to sociodemographic and cognitive ability factors. Controlling for

construct-irrelevant influences may be especially important for performance tasks. Construct underrepresentation is another challenge. Skipping an item on a background questionnaire may be a weak indicator of Conscientiousness, but there are other elements of Conscientiousness not captured with such a measure. It will likely prove fruitful to combine multiple and diverse such indicators to create composite construct measures.

Biodata, Documented Accomplishments, and Administrative Records

Choices made and experiences had influence life directions. Choices and experiences might be gathered from an informant who knows the target, captured with a questionnaire (or personal statement), or reflected on one's resume or bio. Life record data (L-data) was one of three data types identified by Cattell (1965), with T-data and Q-data (questionnaire data). L-data is defined as "behaviour in the actual, everyday life situation" such as "'number of automobile accidents over 20 years,' 'frequency of engagements,' 'number of societies to which the person belongs,' 'marks in school,' and so on" (Cattell, 1965, p. 61). Cattell suggested that someone who knows the person well might be able to provide such data.

Organizational psychology has a long history of research on biodata (biographical data), responses to survey questions about one's own background and experiences. Oswald et al. (2004) included a biodata measure along with the SJT to reflect 12 dimensions of college experience (see the section on SJTs). Some biodata questions were, "How often have you signed a petition for something you believed in? very often/often/sometimes/seldom/never" (citizenship) and "How many times in the last year have you tried to get someone to join an activity in which you were involved or leading? Never/once/twice/three or four times/five times or more" (leadership). Stricker and Rock (1998) reported on an accomplishments questionnaire (designed for graduate school applicants) with items such as "conducted a band, orchestra, or vocal group at a public performance"; "wrote poetry, fiction, or essays that were published"; and "was elected to a major class office in college." These were drawn from dimensions of academic achievement, leadership, language, aesthetics expression, science, and mechanical. Mumford and Owens (1987) and Mumford et al. (2012) reviewed the history, measures, and methodology of biodata, referred to as background data. Biodata measures have been found to be comparable in their predictions of outcomes to Conscientiousness ($r = .30$), although they overlap considerably with cognitive ability, $r = .50$ (Schmidt & Hunter, 1998). Related are behavioral measures of personality (Breil et al., 2022; J. J. Jackson et al., 2010; Soto et al., 2022) based on the specific behaviors individuals engage in, and ambulatory assessments (Trull & Ebner-Priemer, 2013), such as diaries and event sampling.

Also related are work-sample tests for college admissions. The measure is not what one claims on a survey but a sample of actual behavior. Niessen et al. (2016; Meijer & Niessen, 2015; Niessen & Meijer, 2017) provided several suggestions for how such a

behavioral-sampling system could be implemented, such as trial studying approaches (applicant attends a lecture, studies material, takes an exam), and multiple mini-interviews (similar to SJTs) in which applicants are presented with problems they might encounter in professional practice and asked how they would approach them.

As organizations and schools continue to develop increasingly sophisticated and comprehensive data storage systems, such as New Jersey's longitudinal student data warehouse (NJ Smart) or the U.S. National Student Clearinghouse (transcripts and enrollment data), biodata studies can be conducted solely with administrative records without burdening students or teachers. Kautz and Zantoni (2024) combined grades, accumulated credits, absences, and disciplinary infractions, all obtained from school records, to create a noncognitive composite that was as strong a predictor of a variety of future outcomes (11th-grade grades, absences, arrests, graduation, college enrollment, college graduation) as cognitive ability scores were. K. Jackson (2018) combined absences, suspensions, course grades, and grade repetition as a supplement to achievement test scores to create a teacher value-added measure that was more predictive of lagged student outcomes than were test scores alone, which is the typical value-added measure. (Value-added modeling is a controversial approach in educational accountability; see American Statistical Association [2014] and Ho and Polikoff (this volume); the point here is to show that including noncognitive information can potentially improve such measurement. See also C. K. Jackson et al., 2021.)

With advances in data collection technology and statistical methods, big data and machine learning approaches (e.g., regularization techniques to handle wide data [more variables than people]; gradient boosting machines for better out-of-sample prediction), and psychometrics (e.g., latent classes and mixture models to replace biodata's subgrouping methods), there is an opportunity to make advances in biodata measurement. In the early 21st century, such data are much easier to gather through administrative records retained routinely in database systems or on the web. Time spent studying or working on homework is an important determinant of learning (Cooper et al., 2006; Grodner & Rupp, 2013) and is now easier to capture with online learning systems. Variables like registration latency as indicators of procrastination can be captured (Novarese & Di Giovinazzo, 2013). Social media activity relates to personality factors (Kosinski et al., 2013; Youyou et al., 2015), as do contents of offices and workspaces (Gosling et al., 2002). Mobile sensing data from phones, wearables, and beacons differentiate workplace performance levels (Mirjafari et al., 2019). Stress levels can be detected through speech (Slavich et al., 2019). There likely will be rapid developments in these areas.

CONCLUSION AND PROSPECTS

This chapter reviewed noncognitive constructs and methods for assessing them. Noncognitive skills are important determinants and outcomes of education and are in high demand in the workplace. An education benefit is noncognitive in addition to

cognitive skills development, but this benefit is not always apparent because we do not routinely monitor noncognitive skill development. Instead, we infer this relationship because the positive benefits of schooling on a variety of workforce and life outcomes greatly exceeds the benefit attributable to the development of cognitive skills alone. We also know from research involving the administering of noncognitive measures, almost exclusively self- or others reports collected with rating scales, that such measures predict a wide variety of education, workforce, and life outcomes. This relationship is sometimes as strong as and sometimes stronger than the relationship between cognitive measures and those outcomes. However, a recurring finding is that the relationship between noncognitive skills and outcomes is moderated by cognitive ability—the relationship is particularly strong for students of lower cognitive ability and workers. Analyses of the nature of jobs in the workforce, particularly with technology-induced change, suggest that noncognitive skills are likely to increase in importance.

A roadblock to adoption of noncognitive measurement in schools is a widespread but misguided perception that personality is an immutable trait outside the influence of a teacher or the educational system. But ample research demonstrates that personal qualities such as work ethic, social skill, and productive attitudes are as amenable to change due to good teaching and a supportive climate as curricular achievement is. Another barrier has been simple terminology—throughout this article we have used a variety of terms such as personality, social-emotional learning, soft skills, noncognitive skills, interpersonal and intrapersonal skills, attitudes, values, and beliefs almost interchangeably. Each term has a distinct history and constituency but there is a growing recognition that all these terms refer to skills that are fundamentally in the same realm as cognitive skills are. What has largely distinguished noncognitive skills is not the nature of the constructs per se as much as it is the measurement method—cognitive skills tend to be measured with tests and noncognitive skills with ratings. Since 2015 there has been an expansion of frameworks, accompanied by compendia of measures, outlining the nature and extent of noncognitive skills, and we now have a good understanding of what some of the most important skills are.

The literature and practice of noncognitive assessment has been almost completely devoted to the use of rating scale measures, primarily self-ratings. This is due to their ease of development, administration, scoring, and reporting, but there are serious limitations to rating scale measures. They are subject to various unmonitored and uncontrolled biases (reference, response style, social desirability) and the adoption of self-reports limits their use to low-stakes applications. Alternative approaches have benefits—others' reports, forced-choice and ranking approaches, anchoring methods, SJTs, various performance tests of noncognitive skills such as real-effort measures of Conscientiousness, as well as biodata and administrative records, and big data (data mining) approaches. Any measure has advantages and disadvantages in certain situations and it is important to consider the relative strengths and weaknesses of alternatives to determine their suitability for particular applications. Reliability, validity, and fairness analyses can help. Modern psychometric methods—CTT, generalizability theory, IRT (see Cai et al., this volume,

for a discussion of models)—are suitable for almost all noncognitive measurement, even though adoption of proper psychometric methods is not yet widespread. An interpretive-argument perspective on validity is an appropriate filter through which to evaluate noncognitive measurement methods, the data they produce, and the inferences we wish to draw from those data. A major distinction is between low-stakes use, as is currently widely implemented in schools, and high-stakes use, which is growing in workforce personnel selection and increasingly in higher education. Likert-scale self-measurement is inadequate for high-stakes use, but alternatives, such as others' ratings, forced-choice methods, and SJTs, may be adequate for high-stakes use, particularly when treated with appropriate psychometric methods and after pilot testing for susceptibility to coaching and faking. It may also be important to use multiple measurement methods and indicators to mitigate the weaknesses of any one method.

A general theme throughout this chapter has been that the level of interest and value attributed to noncognitive skills in both education and workplace settings, from both the practice and the policy communities, has outpaced the field's ability to provide adequate measures of those skills. We have evidence that interest and perceived importance are likely to grow. It is essential that the educational measurement community meets that interest with appropriate tools.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by funding from the ETS Research Institute, National Science Foundation Award 2201888, and National Institutes of Health Award R01 HD107079. We thank Ida Lawrence, the editors Linda Cook and Mary Pitoniak, and reviewers Fred Oswald, Rob Meijer, and Suzanne Lane, for their inspiring, supportive, constructive, and often extensive comments, which improved the manuscript immeasurably.

REFERENCES

Aicheson, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44*(2)*, 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics, 134*(3), 1121–1162. https://doi.org/10.1093/qje/qjz006

Almlund, M., Duckworth, A., Heckman, J. J., & Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 4, pp. 1–181). Elsevier.

Alreck, P. L., & Settle, R. B. (1994). *The survey research handbook, second edition: Guidelines and strategies for conducting a survey*. McGraw–Hill.

American Association of Medical Colleges. (2020). *AAMC situational judgment test pilot*. https://students-residents.aamc.org/applying-medical-school/preparing-med-school/aamc-situational-judgment-test-pilot/

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

American Statistical Association. (2014, April 8). *ASA statement on using value-added models for educational assessment*. https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Andrich, D. (2016). Rasch rating-scale model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 75–94). Routledge.

Arthur, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99*(3), 535–545. https://doi.org/10.1037/a0035788

Aspen Institute. (2018). *The practice base for how we learn: Supporting students' social, emotional, and academic development. Consensus Statements of Practice from the Council of Distinguished Educators (Sheldon Berman with Sydney Chaffee & Julia Sarmiento)*. https://assets.aspeninstitute.org/content/uploads/2018/03/CDE-Practice-Base_FINAL.pdf

Asplund, J., Agrawal, S., Hodges, T., Harter, J., & Lopez, S. J. (2014). *The Clifton StrengthsFinder® 2.0 technical report: Development and validation*. The Gallup Organization.

Assessment Work Group. (2019). *Student social and emotional competence assessment: The current state of the field and a vision for its future*. Collaborative for Academic, Social, and Emotional Learning.

Atwell, M. N., & Bridgeland, J. M. (2019). *Ready to lead: A 2019 update of principals' perspectives on how social and emotional learning can prepare children and transform schools. A report for CASEL.* Civic with Hart Research Associates.

Autor, D. H., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279–1333.

Autor, D. H., & Price, B. (2013). *The changing task composition of the US labor market: An update of Autor, Levy, and Murnane (2003)*. MIT Economics. https://economics.mit.edu/people/faculty/david-h-autor/working-papers

Bäckströmm, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology, 54*(2), 152–159.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017) Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*, 586–598.

Barnett, S. (1996). *Lives in the balance: Age-27 benefit–cost analysis of the high/scope Perry preschool program* (High/Scope Educational Research Foundation Monograph No. 11). High Scope Press.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barrow, L., & Rouse, C. E. (2005). Do returns to schooling differ by race and ethnicity? *American Economic Review, 95*(2), 83–87. https://doi.org/10.1257/000282805774670130

Barton, D., Farrell, D., & Mourshed, M. (2012). *Education to employment: Designing a system that works*. McKinsey Center for Government. https://www.mckinsey.com/industries/social-sector/our-insights/education-to-employment-designing-a-system-that-works

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263–272.

Bartram, D. (2013). Scalar equivalence of OPQ32: Big Five profiles of 31 countries. *Journal of Cross-Cultural Psychology, 44*(1), 61–83. https://doi.org/10.1177/0022022111430258

Baxter, J. C., Brock, B., Hill, P. C., & Rozelle, R. M. (1981). Letters of recommendation: A question of value. *Journal of Applied Psychology, 66*(3), 296–301. https://doi.org/10.1037/0021-9010.66.3.296

Becker, G. S. (1994). *Human capital: A theoretical and empirical analysis with special reference to education* (3rd ed.). University of Chicago Press.

Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypotheses test *Journal of Educational Measurement, 32*(1), 19-36. https://www.jstor.org/stable/1435190.

Berg, J., Osher, D., Same, M. R., Nolan, E., Benson, D., & Jacobs, N. (2017). *Identifying, defining, and measuring social and emotional competencies—final report*. American Institutes for Research. https://www.air.org/resource/identifying-defining-and-measuring-social-and-emotional-competencies

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*(3), 223–235.

Bertling, J. P. (2015, May 21–24). *Noncognitive modules for the National Assessment of Educational Progress* [Paper presentation]. Association for Psychological Science 27th Annual Convention, New York City, NY, United States.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Addison–Wesley.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I, Cognitive domain*. Longmans, Green and Co.

Blyth, D. A., Borowski, T., Farrington, C. A., Kyllonen, P., & Weissberg, R. P. (2019). *Ten criteria for describing and selecting SEL frameworks*. Collaborative for Academic, Social, and Emotional Learning.

Blyth, D. A., Jones, S., & Borowski, T. (2018). *SEL frameworks—what are they and why are they important?* CASEL. https://measuringsel.casel.org/wp-content/uploads/2018/09/Frameworks-A.1.pdf

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51. https://doi.org/10.1007/BF02291411

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*(4), 453–465. https://doi.org/10.1037/1082-989X.9.4.453

Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behavior Research Methods, 44*, 144–157.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2011). Identification problems in personality psychology. *Personality and Individual Differences, 51*(3), 315–320.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences of the United States of America, 113*(47), 13354–13359.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*, 410–421.

Bowles S., Gintis H., & Osborne M., (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature, 39*(4), 1137–1176. https://doi.org/10.1257/jel.39.4.1137

Bravo-Sanzana, M. V., Varela, J., Terán-Mendoza, O., & Rodriguez-Rivas, M. E. (2023). Measuring school social climate in Latin America: The need for multidimensional and multi-informant tests—A systematic review. *Frontiers in Psychology, (14)*. https://doi.org/10.3389/fpsyg.2023.1190432

Breil, S. M., Mielke, I., Ahrens, H., Geldmacher, T., Sensmeier, J., Marschall, B., & Back, M. D. (2022). Predicting actual social skill expression from personality and skill self-concepts. *Journal of Intelligence, 10*(3), 48. http://dx.doi.org/10.3390/jintelligence10030048

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Broadfoot, A. A. (2006). *Response instructions and faking on situational judgment tests*. Bowling Green State University. http://rave.ohiolink.edu/etdc/view?acc_num=bgsu1161283237

Brown, A. (2016) Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research, 51*(2–3), 345–356. https://doi.org/10.1080/00273171.2016.1150152

Brown, A., & Bartram, D. (2009). *The occupational personality questionnaire revolution: Applying item response theory to questionnaire design and scoring* (Technical report). SHL Group. https://kar.kent.ac.uk/74438/

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate applications series. Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307–333). Routledge.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502.

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52. https://doi.org/10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Bunji, K., & Okada, K. (2019). Item response and response time model for personality assessment via linear ballistic accumulation. *Japanese Journal of Statistics and Data Science, 2*, 263–297. https://doi.org/10.1007/s42081-019-00040-4

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827–854.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O*NET)* (ETS RR-13-21). ETS.

Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental school selection. *Journal of Dental Education, 75*(6), 743–749.

California Department of Education. (2018). *Social and emotional learning in California: A guide to resources*. https://www.cde.ca.gov/eo/in/documents/selresourcesguide.pdf

Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). *Framework for enhancing education and workplace success* (ACT Research Report Series 2015 (4)). ACT. https://files.eric.ed.gov/fulltext/ED558040.pdf

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*, 283–310.

Cao, M. (2016). *Examining the fakability of forced-choice individual differences measures* [Doctoral dissertation]. University of Illinois. http://hdl.handle.net/2142/93064

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics*, Volume 3, Part A (pp. 1801–1863). Elsevier Science.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2020). *The twenty-first mental measurements yearbook*. Buros Center for Testing.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. Partnership for 21st Century Skills.

Castex, G., & Dechter, E. K. (2014). The changing roles of education and ability in wage determination. *Journal of Labor Economics, 32*(4), 685–710.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292–303. https://doi.org/10.1037/h0057299

Cattell, R. B. (1965). *The scientific analysis of personality*. Penguin Books.

Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests*. University of Illinois Press.

Cengage. (2019, January 16). *New survey: Demand for "uniquely human skills" increases even as technology and automation replace some jobs*. https://news.cengage.com/upskilling/new-survey-demand-for-uniquely-human-skills-increases-even-as-technology-and-automation-replace-some-jobs/

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*(1), 143–159. https://doi.org/10.1037/0021-9010.82.1.143

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson, & O. Smit-Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 219–242). Blackwell Publishing.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior and Organization, 149*, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Chernyshenko, O. S. (2003). Applications of ideal point approaches to scale construction and scoring in personality measurement: The development of a six-faceted measure of conscientiousness. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 63*(11-B), 5556.

Chernyshenko, O., Kankaraš, M., & Drasgow, F. (2018). *Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills* (OECD Education Working Papers, No. 173). OECD Publishing. https://doi.org/10.1787/db1d8e59-en

Christian, M. S., Edwards, J. C., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307.

Clark, D., & Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy, 122*(2), 282–318.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs, 14*, 1–56.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*(3), 410–417.

Cochrane, R. C. (1966). *Measures for progress: A history of the National Bureau of Standards*. U.S. Government Printing Office.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Collaborative for Academic, Social, and Emotional Learning. (2020a). *State scan scorecard project*. https://casel.org/state-scan-scorecard-project-2/#info

Collaborative for Academic, Social, and Emotional Learning. (2020b). *SEL assessment guide*. https://measuringsel.casel.org/access-assessment-guide/

Condon, D. M. (2018). *The SAPA Personality Inventory: An empirically derived, hierarchically organized self-report personality assessment model*. PsyArXiv. https://doi.org/10.17605/osf.io/SC4P9

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88*, 322–328.

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.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research, 76*(1), 1–62.

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgment tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention*. John Wiley & Sons. https://doi.org/10.1002/9781118972472.ch11

Costa, A. L., & Kallick, B. (2008). *Learning and leading with habits of mind: 16 essential characteristics for success*. Association for Supervision and Curriculum Development.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Center.

Cox, C. B., Barron, L. G., Davis, W., & de la Garza, B. (2017). Using situational judgment tests (SJTs) in training: Development and evaluation of a structured, low-fidelity scenario-based training method. *Personnel Review, 46*(1), 36–45. https://doi.org/10.1108/PR-05-2015-0137

Credé, M., Tynan, M., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology, 113*, 492–511.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. *International Journal of Selection and Assessment, 19*(4), 363–373.

Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2019). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality and Social Psychology, 117*(3), 674–695. https://doi.org/10.1037/pspp0000210

Darr, W., Borman, W. C., St-Pierre, L., Kubisiak, C., & Grossman, M. (2017). An applied examination of the computerized adaptive rating scale for assessing performance. *International Journal of Selection and Assessment, 25*(2), 149–153.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA Reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502–523.

Debnath, S. C., Lee, B. B., & Tandon, S. (2015). Fifty years and going strong: What makes behaviorally anchored rating scales so perennial as an appraisal method? *International Journal of Business and Social Science, 6*(2), 16–25.

De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P., & Themmen. A. P. N. (2017) Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Science Education, Theory, and Practice, 2*, 243–265. https://doi.org/10.1007/s10459-016-9720-7A

Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics, 132*(4), 1593–1640.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

District of Columbia Public Schools. (2017). *A capital commitment 2017–2022*. https://dcps.dc.gov/capitalcommitment

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(4), 465–476.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support selection and classification decisions* (Technical Report 1311). U.S. Army Research Institute for the Behavioral and Social Sciences.

Duckworth, A. (2013). *Grit: The power of passion and perseverance*. TED Talk Lesson. https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseverance

Duckworth, A. (2016). *Grit: The power of passion and perseverance*. Scribner.

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

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57.

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement, 79*(1), 108–128. https://doi.org/10.1177/0013164417752782

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

Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology, 45*, 294–309.

Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (Eds.). (2015). *Handbook of social and emotional learning: Research and practice* (pp. 3–19). The Guilford Press.

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

Dweck, C. (2014). *The power of believing that you can improve*. TEDxNorrkoping. https://www.ted.com/talks/carol_dweck_the_power_of_believing_that_you_can_improve

Dweck, C. (2017). *Mindset—updated edition: Changing the way you think to fulfill your potential*. Little, Brown Book Group.

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 255–270). Springer.

Ellison, H. B., Grabowski, C. J., Schmude, M., Costa, J. B., Naemi, B., Schmidt, M., Patel, D.; Westervelt, M. (2024). Evaluating a situational judgment test for use in medical school admissions: Two years of AAMC PREview Exam administration data. *Academic Medicine 99*(2), 183–191. 10.1097/ACM.0000000000005548

Elvira, M., & Town, R. (2002). The effects of race and worker productivity on performance evaluations. *Industrial Relations, 40*(4), 571–590.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

ETS. (2014). *ETS standards for quality and fairness*. https://www.ets.org/s/about/pdf/standards.pdf

ETS. (2022). *ETS guidelines for developing fair tests and communications*. https://www.ets.org/pdfs/about/fair-tests-and-communications.pdf

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2015). *The nature and predictive power of preferences: Global evidence.* Institute for the Study of Labor. https://docs.iza.org/dp9504.pdf

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2016). *The preference survey module: A validated instrument for measuring risk, time, and social preferences* (Discussion Paper No. 9674). Institute for the Study of Labor. https://docs.iza.org/dp9674.pdf

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21*(3), 328–347. https://doi.org/10.1037/met0000059

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment, 93*(5), 445–453.

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

Farruggia, S. P., Han, C. W., Watson, L., Moss, T. P., & Bottoms, B. L. (2016). Noncognitive factors and college student success. *Journal of College Student Retention: Research, Theory & Practice, 20*(3), 308–327.

Feng, S., Han, Y., Heckman, J. J., & Kautz, T. (2022). Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills. *Proceedings of the National Academy of Sciences, 119*(6), e2113992119. https://doi.lorg/10.1073/pnas.2113992119

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology, 98*(4), 659–667.

Flake, J. (2017, August 18). Jeff Flake: We need immigrants with skills. But working hard is a skill. *New York Times*. https://www.nytimes.com/2017/08/18/opinion/jeff-flake-we-need-immigrants-with-skills-but-working-hard-is-a-skill.html

Fowler, F. J. (2006). *Survey research methods* (3rd ed.). Sage Publications.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202. https://doi.org/10.1037/0003-066X.39.3.193

French, J. W. (1948). The validity of a persistence test. *Psychometrika, 13*, 271–277. https://doi.org/10.1007/BF02289223

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*, 296–321.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting and Social Change, 114*, 254–280.

Fu, J., Tank, X., & Kyllonen, P. (2022). *Can the generalized graded unfolding model fit dominance statements?* [Unpublished manuscript]. ETS.

Gensowski, M. (2018). Personality, IQ, and lifetime earnings. *Labour Economics, 51*, 170–183. https://doi.org/10.1016/j.labeco.2017.12.004

Gleser, L. J. (1972). On bounds for the average correlation between subtest scores in ipsatively scored tests. *Educational and Psychological Measurement, 32*, 759–765.

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*(2), 251–268. https://doi.org/10.1002/hrm.20278

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34. https://doi.org/10.1037/0003-066x.48.1.26

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. Bantam.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002) A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology, 82*(3), 379–398.

Graham, S. (1991). A review of attribution theory in achievement contexts. *Educational Psychology Review, 3*, 5–39.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Grodner, A., & Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education, 44*(2), 93–109.

Grumm, M., & von Collani, G. (2007). Measuring big-five personality dimensions with the implicit association test—implicit personality traits or self-esteem? *Personality and Individual Differences, 43*(8), 2205–2217. https://doi.org/10.1016/j.paid.2007.06.032

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International*

Journal of Testing, 17(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

Guo, H., Zu, J., & Kyllonen, P. (2019). Consistency of NRM-based scoring rules across cohorts for a situational judgment test. *Psychological Test and Assessment Modeling, 61*, 207–225.

Guo, H., Zu, J., Kyllonen, P., & Schmitt, N. (2016). *Evaluation of different scoring rules for a non-cognitive test in development* (ETS Research Report RR-16-03). ETS.

Guttman, L., & Levy, S. (1982). On the definition and varieties of attitude and wellbeing. *Social Indicators Research, 10*, 159–174.

Gwet, K. L. (2014). Handbook of inter-rater reliability (4th ed.). Advanced Analytics Press.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*, 79–95.

Hall, E. P., Gott, S. P., & Pokorny, R. A. (1995). *A procedural guide to cognitive task analysis: The PARI methodology* (Technical Report ADA303654). Armstrong Lab, Brooks Air Force Base. https://apps.dtic.mil/dtic/tr/fulltext/u2/a303654.pdf

Hamilton, L. S., Stecher, B. M., Schweig, J., & Baker, G. (2018). *RAND Education Assessment Finder*. RAND Corporation. https://www.rand.org/pubs/tools/TL308.html

Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. C. (2017). Initial steps towards a standardized assessment for CPS: Practical challenges and strategies. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 135–156). Springer.

Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. (2019). *Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving.* [ETS Research Report No. RR-19-41]. ETS. https://doi.org/10.1002/ets2.12276.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*(1), 43–62. https://doi.org/10.1111/j.1744-6570.1988.tb00631.x

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*(3), 319–334.

He, J., & van de Vijver, F. J. (2015). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly, 79*, 267–290.

Heckman, J. J., Humphries, J. E., & Kautz, T. (Eds.). (2014). *The myth of achievement tests: The GED and the role of character in American life*. University of Chicago Press.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review, 91*(2), 145–149. https://doi.org/10.1257/aer.91.2.145

Hedlund, J., Wilt, J. M., Nebel, K. R., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the Graduate Management Admissions Test. *Learning and Individual Differences, 16*, 101–127.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.

Hilton, M. (2019). Education for workforce readiness: Findings from reports of the National Academies of Sciences, Engineering, and Medicine. In F. L. Oswald, T. S. Behrens, & L. L. Foster (Eds.), *Workforce readiness and the future of work* (pp. 189–206). Routledge.

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

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement, 39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Hontangas, P. M., Leenen, I., de la Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema, 28*(1), 76–82.

Horn, J. L. (1971). Motivation and dynamic calculus concepts from multivariate experiment. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (2nd printing, pp. 611–641). Rand McNally.

Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement*. Policy Analysis for California Education. https://edpolicyinca.org/sites/default/files/SEL-CC_report.pdf

Inchley, J., Currie, D., Cosma, A., & Samdal, O. (Eds.). (2018). *Health behaviour in school-aged children (HBSC) study protocol: Background, methodology and mandatory items for the 2017/18 survey*. CAHRU.

Institute of Educational Sciences. (2007). *Character education: WWC evidence review protocol for character education interventions*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/CharEd_protocol.pdf https://eric.ed.gov/?id=ED497054

Institute of Educational Sciences. (2020). *WWC What Works Clearinghouse*. U.S. Department of Education. https://ies.ed.gov/ncee/wwc/

Ion, A., Gunnesch-Luca, G., Petre, D., & Iliescu, D. (2022). Secular changes in personality: An age-period-cohort analysis. *Journal of Research in Personality, 100*, 104280. https://doi.org/10.1016/j.jrp.2022.104280

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2021). Linking social-emotional learning to long-term success: Student survey responses show effects in high school and beyond. *Education Next, 21*(1), 64–71.

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2016). The internal structure of situational judgement tests reflects candidate main effects:

Not dimensions or situations. *Journal of Occupational and Organizational Psychology, 90*(1), 1–27.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*(4), 243–252. https://doi.org/10.1037/h0045996

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality, 44*(4), 501–511. https://doi.org/10.1016/j.jrp.2022.104280

Jackson, K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy, 126*(5), 2072–2107.

James, L. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*(2), 131–163. https://doi.org/10.1177/109442819812001

John, O. P., & De Fruyt, F. (2015). *Education and social progress: Framework for the longitudinal study of social and emotional skills in cities* (EDU-CERI-CD(2015)13.en). OECD Publishing. https://one.oecd.org/document/EDU/CERI/CD(2015)13/en/pdf

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Jones, S., Bailey, R., Brush, K., & Nelson, B. (2019). *Introduction to the taxonomy project: Tools for selecting & aligning frameworks*. CASEL. https://measuringsel.casel.org/wp-content/uploads/2019/02/Frameworks-C.1.pdf

Judge, T. A., & Higgins, C. (1998). Affective disposition and the letter of reference. *Organizational Behavior and Human Decision Processes, 75*, 207–221.

Kane, M., Berryman, S., Goslin, D., & Meltzer, A. (1990). *The Secretary's Commission on Achieving Necessary Skills: Identifying and describing the skills required by work.* Pelavin Associates.

Kankaraš, M. (2017). *Personality matters: Relevance and assessment of personality characteristics* (OECD Education Working Paper No. 157). OECD Publishing.

Kankaraš, M., & Suarez-Alvarez, J. (2019). *Assessment framework of the OECD Study on Social and Emotional Skills* (OECD Education Working Papers 207). OECD Publishing. https://doi.org/10.1787/5007adef-en

Kapteyn, A., Smith, J. P., van Soest, A. H. O., & Vonkova, H. (2011). *Anchoring vignettes and response consistency* (RAND Working Paper Series WR-840). RAND. https://ssrn.com/abstract=1799563 or http://dx.doi.org/10.2139/ssrn.1799563

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of Situational Judgment Tests (SJTs). *European Journal of Psychological Assessment, 32*(3), 230–240.

Kasten, N., Freund, P. A., & Staufenbiel, T. (2018). "Sweet little lies": An in-depth analysis of faking behavior on situational judgment tests compared to personality questionnaires. *European Journal of Psychological Assessment, 36*(1), 136–148. http://dx.doi.org/10.1027/1015-5759/a000479

Kautz, T., & Zantoni, W. (2024). Measurement and development of noncognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal program. *Journal of Human Capital, 18*(2), 272–304. DOI: 10.1086/728087

Keener, A. (2020). *A comparison of Cohen's Kappa and Gwet's AC1 with a mass shooting classification index: A study of rater uncertainty* [Doctoral dissertation]. Oklahoma State University, Stillwater.

Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P. A., Chen, L., & Feng, G. (2017). *Exploring methods for developing behaviorally anchored rating scales for evaluating structured interview performance* (ETS Research Report RR-17-28). ETS.

Khorramdel, L., Jeon, M., & Wang, L. L. (2019). Editorial: Advances in modeling response styles and related phenomena. *British Journal of Mathematical and Statistical Psychology, 72*(3), 393–400.

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49*(2), 161–177. https://doi.org/10.1080/00273171.2013.866536

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15*(1), 46–66.

Klieger, D. M., Kell, H. J., Rikoon, S., Burkander, K. N., Bochenek, J. L., & Shore, J. R. (2018). *Development of the behaviorally anchored rating scales for the skills demonstration and progression guide* (ETS Research Report RR-18-24). ETS.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kubinger, K. D. (2009). The technique of objective personality-tests sensu R. B. Cattell nowadays: The Viennese pool of computerized tests aimed at experiment-based assessment of behavior. *Acta Psychological Sinica, 41*, 1024–1036.

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment, 22*, 101–107. https://doi.org/10.1111/ijsa.12060

Kuncel N. R., & Tellegen A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*, 201–228.

Kyllonen, P. C. (2006). *The research behind the ETS® Personal Potential Index. A background paper from ETS*. ETS. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.522.5309&rep=rep1&type=pdf

Kyllonen, P. C. (2013). Soft skills for the workplace. *Change. The Magazine for Higher Learning, 45*(6), 16–23.

Kyllonen, P. C. (2015). *Developing definitions and assessment methods. National Academies of Sciences, Engineering, and Medicine* [Presentation]. Symposium on assessing hard-to-measure cognitive, interpersonal, and intrapersonal competencies, Washington, DC, United States.

Kyllonen, P. C. (2016). Socio-emotional and self-management variables in learning and assessment. In A. Rupp & J. Leighton (Eds.), *Handbook of cognition and assessment* (pp.174–197). Wiley/Blackwell.

Kyllonen, P. C., Heincke, P., Holtzman, S., Olivera-Aguilar, M., Iacovelli, M., Xu, J., Grodman, L., & DelMonico, B. (2020). *Predicting Yale School of Management outcomes with a forced-choice behavioral assessment* [Unpublished manuscript]. ETS.

Kyllonen, P. C., & Bertling, J. P. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). CRC Press.

Kyllonen, P. C., & Kell, H. (2018). Ability tests measure personality, personality tests measure ability: Disentangling construct and method in evaluating the relationship between personality and ability. *Journal of Intelligence, 6*(3), 32. https://doi.org/10.3390/jintelligence6030032

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86*(2), 602–640. https://doi.org/10.3102/0034654315617832

LeBreton, J. M., Grimaldi, E. M., & Schoen, J. L. (2020). Conditional reasoning: A review and suggestions for future test development and validation. *Organizational Research Methods, 23*(1), 65–95. https://doi.org/10.1177/1094428118816366

Legree, P. J., Kerner, B. S., & Shewach, O. R. (2021). *Identifying optimal keys to enhance personality scale validity* (Research Note 2021-04). U.S. Army Research Institute for the Behavioral and Social Sciences.

Lee, J., & Stankov, L. (2013). Higher-order structure of non-cognitive constructs and prediction of PISA 2003 math achievement. *Learning and Individual Differences, 26*, 119–130. https://doi.org/10.1016/j.lindif.2013.05.004

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lee, P., Joo, S.-H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement, 43*(3), 226–240. https://doi.org/10.1177/0146621618768294

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education, 2*(1), 1–24.

Levin, H. M. (1989). Ability testing for job selection: Are the economic claims justified? In B. R. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 211–232). National Commission on Testing and Public Policy.

Levin, H. M. (2012). More than just test scores. *Prospects: Quarterly Review of Comparative Education, 42*, 269–284.

Likert, R. (1932). A technique for measurement of attitudes. *Archives of Psychology, 140*, 5–55.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486–512.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101–128.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94*(4), 1095–1101.

Lombardo, M. M., & Eichinger, R. W. (2004). *The leadership machine* (3rd ed.). Lominger International.

Lubke, G. H., & Muthen, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*, 514–534.

Ludlow, L., Anghel, E., Szendey, O., O'Keefe, T., Howell, B., Matz-Costa, C., & Braun, H. (2020). The Boston College Living a Life of Meaning and Purpose (BC-LAMP) portfolio: An application of Rasch/Guttman scenario methodology. *Journal of Applied Measurement, 21*(2). 134–153.

Ludlow, L. H., Matz-Costa, C., & Klein, K. (2019). Enhancement and validation of the Productive Engagement Portfolio–Scenario (PEP–S8) Scales. *Measurement and Evaluation in Counseling and Development, 52*(1), 15–37. https://doi.org/10.1080/07481756.2018.1497430

Madnani, N., & Cahill, A. (2018). *Automated scoring: Beyond natural language processing*. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics (pp. 1099–1109). Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1094.pdf

Mahoney, J. L., Durlak, J. A., & Weissberg, R. P. (2018). An update on social and emotional learning outcome research. *Phi Delta Kappan, 100*(4), 18–23.

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S. (2017). *Jobs lost, job gained: What the future of work will mean for jobs, skills, and wages*. McKinsey & Company. https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*, 311–360.

Martin-Raugh, M. P., Kyllonen, P. C., Hao, J., Bacall, A., Becker, D., Kurzum, C., Yang, Z., Yan, F., & Barnwell, P. (2020). Negotiation as an interpersonal skill: Generalizability

of negotiation outcomes and tactics across contexts at the individual and collective levels. *Computers in Human Behavior, 104*, 105966. https://doi.org/10.1016/j.chb.2019.03.030

Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 109–128). Routledge.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*(6), 935–974. https://doi.org/10.1080/00273171.2010.531231

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review, 123*(5), 569–591. https://doi.org/10.1037/rev0000035

McAbee, S. T., & Oswald, F. L. (2013). The criterion-related validity of personality measures for predicting GPA: A meta-analytic validity competition. *Psychological Assessment, 25*(2), 532–544.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability: Monograph*. Rand Corporation.

McCaffrey, D. F., Oliveri, M. E., & Holtzman, S. (2018). *A generalizability theory study to examine sources of score variance in third-party evaluations used in decision-making for graduate school admissions* (GRE Board Research Report No. GRE-18-03). ETS. https://doi.org/10.1002/ets2.12225

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730–740.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96*(2), 327–336.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*, 515–525. https://doi.org/10.1016/j.intell.2005.02.001

Medvedev, O. N., Merry, A. F., Skilton, C., Gargiulo, D. A., Mitchell, S. J., & Weller, J. M. (2019). Examining reliability of WHOBARS: A tool to measure the quality of administration of WHO surgical safety checklist using generalizability theory with surgical teams from three New Zealand hospitals. *BMJ Open, 9*(1), e022625. https://doi.org/10.1136/bmjopen-2018-022625

Meijer, R. R., & Niessen, A. S. M. (2015). A trial studying approach to predict college achievement. *Frontiers in Psychology, 6*, 887.

Messick, S. (1978). *Potential uses of noncognitive measurement in education* (ETS Research Bulletin Series, 1, 1–25). ETS. https://doi.org/10.1002/j.2333-8504.1978.tb01156.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Meyer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology, 59*, 507–536. https://doi.org/10.1146/annurev.psych.59.103006.093646

Meyer, R. H., Wang, C., & Rice, A. B. (2018). *Measuring students' social-emotional learning among California's CORE Districts: An IRT modeling approach* [Working paper]. Policy Analysis for California Education. https://www.edpolicyinca.org/sites/default/files/Measuring_SEL_May-2018.pdf

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Ministerio de Educacion Goblerno de Chile [Chilean Government Ministry of Education]. (2016). *Plan de evaluaciones nacionales e internacionales para el periodo 2016–2020* [National and international evaluation plan for the period 2016–2020]. Division Juridica, Solicitud No. 00197, Decreto No. 01820 (20 June 2016).

Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P. G., Campbell, A. T., Chawla, N. V., Das Swain, V., De Choudhury, M., Dey, A. K., D'Mello, S. K., Gao, G., Gregg, J. M., Jagannath, K., Jiang, K., Lin, S., Liu, Q., Mark, G., Martínez, G. J., . . . Striegel, A. (2019). Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(2), 1–24. ACM. https://dl.acm.org/doi/10.1145/3328908

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity, 42*(6), 779–794. https://doi.org/10.1007/s11135-006-9067-x

Moretti, E. (2005, March 1). Social returns to human capital. *The Reporter*, The National Bureau of Economic Research. https://www.nber.org/reporter/spring-2005/social-returns-human-capital

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 40*(7), 500–516. https://doi.org/10.1177/0146621616662226

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G., Cabrera, H. F., Chen, S. X., Church, A. T., Cissé, D. D., . . . Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*, 1423–1436. https://doi.org/10.1177/0146167212451275

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*(2), 321–333. http://dx.doi.org/10.1037/a0017975

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*(1), 1–31. https://doi.org/10.1177/014662168701100101

Mumford, M. D., Whetzel, D. L., Murphy, S. T., & Eubanks, D. L. (2012). Background data. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 201–233). Routledge.

Muraki, E., & Muraki, M. (2016). Generalized partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 127–138). Routledge.

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in academic admissions: A meta-analysis and cautionary tale. *College and University, 84*(4), 83–88.

Muthén, L. K., & Muthén, B. O. (1998–2011). *Mplus User's Guide* (6th ed.). Muthén & Muthén.

Naemi, B. D., Seybert, J., Robbins, S. B., & Kyllonen, P. C. (2014). *Examining the WorkFORCE™ Assessment for Job Fit and Core Capabilities of FACETS™* (ETS Research Report RR-14-32). https://doi.org/10.1002/ets2.12040

National Academies of Sciences, Engineering, and Medicine. (2017). *Supporting students' college success: The role of assessment of intrapersonal and interpersonal competencies*. The National Academies Press. https://doi.org/10.17226/24697

National Assessment Governing Board. (2002). *Background information framework for the National Assessment of Educational Progress*.

National Assessment Governing Board. (2013). *Contextual information framework for the National Assessment of Educational Progress*. https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/contextual-information/contextual-information-framework.pdf

National Association of Colleges and Employers. (2018). *2019 NACE job outlook report.* https://www.odu.edu/content/dam/odu/offices/cmc/docs/nace/2019-nace-job-outlook-survey.pdf

National Center for Education Statistics. (n.d.-a). *National Education Longitudinal Study of 1988 (NELS:88): Questionnaires*. Institute of Education Sciences. https://nces.ed.gov/surveys/nels88/questionnaires.asp

National Center for Education Statistics. (n.d.-b). *Survey questionnaires: What can survey questionnaires tell us about student achievement?* Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/experience/survey_questionnaires.aspx

National Research Council. (2008). *Research on future skill demands: A workshop summary*. The National Academies Press. https://doi.org/10.17226/12066.

National Research Council (2011). *Assessing 21st century skills: Summary of a workshop*. The National Academies Press. https://doi.org/10.17226/13215

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398

Nering, M., & Ostini, R. (Eds.) (2010). *Handbook of polytomous item response theory models*. Routledge.

Neubauer, A. C., & Hofer, G. (2022). (Retest-)reliable and valid despite low alphas? An example from a typical performance situational judgment test of emotional

management. *Personality and Individual Differences, 189*, 111511. https://doi.org/10.1016/j.paid.2022.111511

Newstead, S. E., & Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. *Ergonomics, 30*(10), 1447–1462. https://doi.org/10.1080/00140138708966038

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*(4), 250–260. https://doi.org/10.1111/j.1468-2389.2005.00322.x

Niessen, A. S. M., & Meijer, R. R. (2017). On the use of broadened admission criteria in higher education. *Perspectives on Psychological Science, 12*(3), 436–448. https://doi.org/10.1177/1745691616683050

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Predicting success in higher education using proximal predictors. *Plos ONE, 11*(4), e0153663.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences, 106*, 183–189.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93*(1), 116–130.

Novarese, M., & Di Giovinazzo, V. (2013). *Promptness and academic performance* (MPRA Paper No. 49746). University Library of Munich, Germany. http://mpra.ub.uni-muenchen.de/49746/

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*(5), 966–980. https://doi.org/10.1037/a0022955

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L. III, & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*(1), 19–29. https://doi.org/10.1111/j.1468-2389.2007.00364.x

O'Toole, B. I., & Stankov, L. (1992). Ultimate validity of psychological tests. *Personality and Individual Differences, 13*(6), 699–716.

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*(4), 762–773.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Pomeranz, H. R., & Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation. *Academy of Management Learning & Education, 11*(4), 609–630. http://dx.doi.org/10.5465/amle.2010.0177

Oostrom, J. K., De Soete, B., & Lievens, F. (2015). Situational judgment testing: A review and some new developments. In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 172–189). Routledge.

Oostrom, J. K., Kobis, N. C., Ronay, R., & Cremers, M. (2017). False consensus in situational judgment tests: What would others do? *Journal of Research in Personality, 71*, 33–45.

Organisation for Economic Co-operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 Technical Report*. https://www.oecd.org/en/publications/pisa-2012-technical-report_6341a959-en.html

Organisation for Economic Co-operation and Development. (2015). *Skills for social progress: The power of social and emotional skills*. OECD Publishing. http://dx.doi.org/10.1787/9789264226159-en

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 assessment and analytical framework*: *Science, reading, mathematic and financial literacy*. OECD Publishing. https://doi.org/10.1787/9789264255425-1-en

Organisation for Economic Co-operation and Development. (2017a). *PISA 2015 database*. https://www.oecd.org/pisa/data/2015database/

Organisation for Economic Co-operation and Development. (2017b). PISA 2012 Technical Report. https://www.oecd.org/en/publications/pisa-2012-technical-report_6341a959-en.html

Organisation for Economic Co-operation and Development. (2019a). *OECD study on social and emotional skills*. https://www.oecd.org/education/ceri/OECD-Study-on-Social-and-Emotional-Skills.pdf

Organisation for Economic Co-operation and Development. (2019b). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Organisation for Economic Co-operation and Development. (2021). *Beyond academic learning: First results from the Survey of Social and Emotional Skills*. OECD Publishing https://doi.org/10.1787/92a11084-en

Organisation for Economic Co-operation and Development. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en.html

Oswald, F. L., Behrend, T. S., & Foster, L. (Eds.). (2019). *Workforce readiness and the future of work*. Routledge.

Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right—but so far, Likert was not wrong. *Industrial and Organizational Psychology, 3*(4), 481–484.

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. http://dx.doi.org/10.1037/0021-9010.89.2.187

Ortner, T., & Proyer, R. (2015). Objective personality tests (pp. 133–149). In T. M. Ortner & F. J. R. van de Vijver (Eds.), *Behavior-based assessment in psychology*. Hogrefe & Huber.

Ortner, T. M., & Schmitt, M. (2014). Advances and continuing challenges in objective personality testing. *European Journal of Psychological Assessment, 30*(3), 163–168. http://dx.doi.org/10.1027/1015-5759/a000213

Ozer, D., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401–421.

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice. *AMEE Guide No. 100, 38*(1), 3–17. https://doi.org/10.3109/0142159X.2015.1072619

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 67–88). Lawrence Erlbaum Associates.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*(3), 524–539. https://doi.org/10.1037/0022-3514.81.3.524

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences, 183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*(1), 70–89.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook of classification*. American Psychological Association; Oxford University Press.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET*. American Psychological Association. https://doi.org/10.1037/10313-000

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*(1), 1–16. https://doi.org/10.1111/1468-2389.00222

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7*, 27–65. http://dx.doi.org/10.1177/1094428103259554

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996

Poropat, A. (2014). Other-rated personality and academic performance: Evidence and implication. *Learning and Individual Differences, 34*, 24–32.

Prasad, J. (2017). *It's both who you are and where you're from: Relating vocational interests and socioeconomic status to bias in biodata and SJTs* [Master's thesis, Michigan State University]. https://d.lib.msu.edu/etd/6934

Prasad, J. J., Showler, M. B., Ryan, A. M., Schmitt, N., & Nye, C. D. (2017). When belief precedes being: How attitudes and motivation before matriculation lead to fit and academic performance. *Journal of Vocational Behavior, 100*, 27–42. https://doi.org/10.1016/j.jvb.2017.02.003

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*(3), 523–550

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment, 32*, 39–51. https://doi.org/10.1027/1015-5759/a000336

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology, 72*(3), 447–465. https://doi.org/10.1111/bmsp.12168

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling, 55*(4), 361–382.

Rath, T. (2007). *StrengthsFinder 2.0*. Gallup Press.

Rath, T., & Conchie, B. (2008). *Strengths based leadership: Great leaders, teams, and why people follow*. Gallup Press.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Reeves, R. V., & Venator, J. (2014). *Jingle-jangle fallacies for non-cognitive factors*. Brookings. https://www.brookings.edu/articles/jingle-jangle-fallacies-for-non-cognitive-factors/

Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research, 43*(1), 73–97. https://doi.org/10.1177/0049124113509605

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353–387.

Rieger, S., Göllner, R., Spengler, M., Trautwein, U., Nagengast, B., & Roberts, B. W. (2017). Social cognitive constructs are just as stable as the Big Five between grades 5 and 8. *AERA Open, 3*(3). https://doi.org/10.1177/2332858417717691

Rimfeld, K., Malanchini, M., Krapohl, E., Hannigan, L. J., Dale, P. S., & Plomin, R. (2018). The stability of educational achievement across school years is largely explained by genetic factors. *Science of Learning*, *16*(3), 1–10. https://doi.org/0.1038/s41539-018-0030-0

Rios, J. A., Ling, G., Pugh, R., Becker, D. M., & Bacall, A. N. (2020). Identifying critical 21st century for workplace success: A content analysis of job advertisements. *Educational Researcher, 49*, 80–89.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261–288.

Robbins, S. B., Oh, I.-S., Le, H., & Button, C. (2009). Intervention effects on college performance and persistence, mediated by motivational, emotional, and social control factors: Integrated meta-analytic path analyses. *Journal of Applied Psychology, 94*, 1163–1184.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality, 43*, 137–145. https://doi.org/10.1037/0033-2909.132.1.1

Roberts, B. W., Kuncel, N., Shiner, R. N., Caspi, A., & Goldberg, L. (2007). The power of personality: A comparative analysis of the predictive validity of personality traits, SES, and IQ. *Perspectives in Psychological Science, 2*, 313–345.

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin, 143*(2), 117–141. https://doi.org/0.1037/bul0000088

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3–32. https://doi.org/10.1177/01466216000241001

Rothmann, S., & Coetzer, E. P. (2003). The Big Five personality dimensions and job performance. *Journal of Industrial Psychology, 29*, 68–74.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80*(1), 1-28. https://doi.org/10.1037/h0092976

Salgado, J. F. (1997). The five-factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82, 30–43*.

Salgado, J. F., & Lado, M. (2018). Faking resistance of a quasi-ipsative forced-choice personality inventory without algebraic dependence. *Journal of Work and Organizational Psychology, 34*(3), 213–216. https://doi.org/10.5093/jwop2018a23

Salgado, J. F., & Tauriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3–30. https://doi.org/10.1080/1359432X.2012.716198

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups:

A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88*(4), 797–834.

Salzburg Global Seminar. (2018). *The Salzburg statement for social and emotional learning*. https://www.salzburgglobal.org/fileadmin/user_upload/Documents/2010-2019/2018/Session_603/SalzburgGlobal_Statement_SEL.pdf

Samejima, F. (2016). Graded response model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 77–108). Routledge.

Sass, R., Frick, S., Reip, U.-D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivational forced-choice versus rating scale instruments. *Assessment, 27*(3), 572–584. https://doi.org/10.1177/1073191118762049

Saucier, G. (2000). Isms and the structure of social attitudes. *Journal of Personality and Social Psychology, 78*(2), 366–385. https://doi.org/10.1037/0022-3514.78.2.366

Saucier, G. (2013). Isms dimensions: Toward a more comprehensive and integrative model of belief-system components. *Journal of Personality and Social Psychology, 104*, 921–939.

Schaepers, P. C., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, K. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality, 87*, 1–5.

Schanzenbach, D. W., Nunn, R., Bauer, L., Mumford, M., & Breitwieser, A. (2016). *Seven facts on noncognitive skills from education to the labor market.* The Hamilton Project: https://www.hamiltonproject.org/assets/files/seven_facts_noncognitive_skills_education_labor_market.pdf

Schläpfer, F., & Fischhoff, B. (2010). *When are preferences consistent? The effects of task familiarity and contextual cues on revealed and stated preferences* (Working Paper No. 1007). University of Zurich, Socioeconomic Institute. https://www.econstor.eu/bitstream/10419/76190/1/638617693.pdf

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

Schmitt, N., & Ali, A. A. (2015). The practical importance of measurement invariance. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 327–346). Routledge.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology, 94*(6), 1479–1497.

Schunk, D. H. (1989). Self-efficacy and achievement behaviors. *Educational Psychology Review, 1*, 173–208.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438–1457.

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association, 11(4)*, 743–779. https://doi.org/10.1111/jeea.12025

Seybert, J., & Becker, D. (2019). *Examination of the test-retest reliability of a forced-choice personality measure* (ETS Research Report RR-19-37). ETS.

SHL Group. (1999). *OPQ32: Manual and user's guide*.

Shultz, M. M., & Zedeck, S. (2011). Predicting lawyering effectiveness: Broadening the basis for law school admission decisions. *Law & Social Enquiry, 36*(3), 620–661. https://doi.org/10.1111/j.1747-4469.2011.01245.x

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science, 29*(4), 549–571. https://doi.org/10.1177/0956797617739704

Slavich, G. M., Taylor, S., & Picard, R. W. (2019). *Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations*. *Stress, 22*(4), 408–413. https://doi.org/10.1080/10253890.2019.1584180

Society for Human Resource Management. (2014). *SHRM Body of Competency and Knowledge*™. https://snv.shrm.org/sites/snv.shrm.org/files/SHRMBoCK.pdf

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st century competencies: Guidance for educators*. Asia Society. https://asiasociety.org/files/gcen-measuring21cskills.pdf

Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: *The BESSI Journal of Personality and Social Psychology, 123*(1), 192–222. https://doi.org/10.1037/pspp0000401

Stankov, L., Morony, S., & Lee, Y.-P. (2013). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology, 34*(1), 9–28. https://doi.org/10.1080/01443410.2013.814194

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184–203.

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

Stemler, S. E., Bebell, D., & Sonnabend, L. A. (2011). Using school mission statements for reflection and research. *Educational Administration Quarterly, 47*(2), 383–420. https://doi.org/10.1177/0013161X10387590

Sternberg, R. J., & Horvath, J. A. (1999). *Tacit knowledge in professional practice: Researcher and practitioner perspectives*. Lawrence Erlbaum Associates.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. Cambridge University Press.

Stone, T. H., Foster, J., Webster, B. D., Harrison, J., & Jawahar, I. M. (2016). Gender differences in supervisors' multidimensional performance ratings: Large sample evidence. *Human Performance, 29*(5), 428–446. http://dx.doi.org/10.1080/08959285.2016.1224884

Stricker, L. J., & Rock, D. A. (1998). Assessing leadership potential with a biographical measure of personality traits. *International Journal of Selection and Assessment, 6*(3), 164–184. https://doi.org/10.1111/1468-2389.00087

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling, 27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Taylor, R. D., Oberle, E., Durlak, J. A., Weissberg, R. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development, 88*(4), 1156–1171.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703–742.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Wiley.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*(1), 151–176.

U.S. Office of Personnel Management. (2018a). *Ssituational judgment tests*. https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/situational-judgment-tests/

U.S. Office of Personnel Management. (2018b). *Assessment & selection*. https://www.opm.gov/policy-data-oversight/assessment-and-selection/

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (Ed.). (2016). Handbook of item response theory. Chapman and Hall/CRC.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality & Social Psychology, 98*(2), 281–300.

Vedel, A. (2014). The Big Five and tertiary academic performance: A systematic review and meta-analysis. *Personality and Individual Differences, 71*, 66–76. https://doi.org/0.1016/j.paid.2014.07.011

von Davier, M. (2016). Rasch model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 31–50). Routledge.

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement, 42*(4), 291–306. https://doi.org/10.1177/0146621617730389

Vukasovic, T., & Bratko, D. (2015). Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin, 141*(4), 769–785. https://doi.org/10.1037/bul0000017

Wand, J., King, G., & Lau. O. (2011). Anchors: Software for anchoring vignettes data. *Journal of Statistical Software, 42*(3), 1–25.

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement, 41*(8), 600–613. https://doi.org/10.1177/0146621617703183

Wanzer, D., Postlewaite, E., & Zargarpour, N. (2019). Relationships among noncognitive factors and academic performance: Testing the University of Chicago Consortium on School Research Model. *AERA Open, 5*(4), 1–20. https://doi.org/10.1177/2332858419897275

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science, 29*(7), 1159–1177. https://doi.org/10.1177/0956797618761661

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *Review of Economics and Statistics, 96*(4), 849–861.

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

Weng, Q., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior, 104*, 199–209.

West, M. R. (2016). *Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts*. Brookings. https://www.brookings.edu/research/should-non-cognitive-skills-be-included-in-school-accountability-systems-preliminary-evidence-from-californias-core-districts/

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE districts. *Journal of Applied Developmental Psychology, 55*, 119–129. https://doi.org/10.1016/j.appdev.2017.06.001

West, M. R., Pier, L., Fricke, H., Loeb, S., Meyer, R. H., & Rice, A. B. (2018).*Trends in student social-emotional learning: Evidence from the CORE Districts* (Working paper). PACE: Policy Analysis for California Education, CORE-PACE Research Partnership. https://edpolicyinca.org/publications/trends-student-social-emotional-learning

Westring, A. J. F., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance, 22*(1), 44–63. https://doi.org/10.1080/08959280802540999

Whelpley, C. E. (2014). *How to score situational judgment tests: A theoretical approach and empirical test* [Doctoral dissertation, Virginia Commonwealth University]. https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=4607&context=etd

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3), 188–202. https://doi.org/10.1016/j.hrmr.2009.03.007

Wise, S. L., & Gao, L. (2014). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality, 19*, 373–390.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*, 364–369. https://doi.org/10.1038/s41586-019-1466-y

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zamarro, G., Cheng, A., Shakeel, M. D., & Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics, 72*(C), 51–60. https://doi.org/10.1016/j.socec.2017.11.005

Zu, J., & Kyllonen, P. C. (2020). Nominal response model is useful for scoring multiple-choice situational judgment tests. *Organizational Research Methods, 23*(2), 342–366. https://doi.org/10.1177/1094428118812669

NOTES

1. Selection due to death could explain some of these findings, as it has with cognitive ability (O'Toole & Stankov, 1992).
2. Meta-analyses typically report both a raw Spearman correlation (r) and a disattenuated correlation adjusted for measurement error in the predictor (ρ).
3. Alternatively, standardized alpha is based on average interitem correlation rather than covariance. C. F. Falk and Savalei (2011) discussed the differences in meaning between the two alphas with respect to personality measures. If a composite is the sum of the raw item scores because items are on the same scale, alpha is justified; if a composite is the sum of standardized item scores, as a result of items being on different scales, standardized alpha is justified.
4. Other more general IRT treatments can be found in Embretson and Reise (2000) and de Ayala (2009). Specialized treatments covering polytomous (multiple-category) IRT models can be found in Thissen and Wainer (2001) and Nering and Ostini (2010).