

Assessing Non-Cognitive Constructs in Education: A Review of Traditional and Innovative Approaches

Anastasiya A. Lipnevich, Carolyn MacCann, and Richard D. Roberts

Abstract

This chapter provides a broad overview of both conventional and novel approaches for assessing non-cognitive skills, specifically focusing on their application in educational contexts. Conventional approaches include self-assessments, other-ratings, letters of recommendation, biodata, and interviews. We outline the current uses and validity evidence for these methods, and discuss the theory of planned behavior as a useful heuristic for assessment development. Novel approaches to non-cognitive assessment include the situational judgement test, day-reconstruction method, and use of writing samples. After reviewing these new approaches, we discuss the issue of response distortion in non-cognitive assessment, outlining some assessment techniques thought to be less susceptible to faking. Suggested fake-resistant assessments include the implicit association test and conditional reasoning test, as well as forced-choice tests and the Bayesian truth serum. We conclude with a series of summary statements concerning uses of non-cognitive testing in education.

Key Words: non-cognitive assessment, self-report, other-reports, situational judgement test, biodata

Introduction

It is increasingly obvious that succeeding in education requires much more than just "book smarts." To succeed, students need to be not just intelligent, they need to work hard, believe in themselves, cope with the stress of academic evaluations, develop and maintain networks of social and academic support, and organize their homework, projects, and study. That is, students' non-cognitive qualities can be as influential as their cognitive skills in influencing their academic achievement and educational aspirations (e.g., Burrus, MacCann, Kyllonen, & Roberts, 2011). Moreover, these non-cognitive qualities do not just predict the grades awarded by teachers or schools, but the hard data collected by large-scale testing programs. For example, research demonstrates that children and adolescent's levels of self-efficacy, self-concept, and self-confidence predict their mathematics, science, and reading scores

(Campbell, Voelkl, & Donahue, 1997; Connell, Spencer, & Aber, 1994). Non-cognitive constructs are important predictors of academic achievement and behavioral adjustment from early childhood, with Abe's (2005) research demonstrating that personality at age three predicted academic achievement in later schooling.

Perhaps even more importantly, these non-cognitive constructs are not simply proxies for a privileged background or for student ability variables. Duckworth and Seligman (2005) demonstrate that non-cognitive variables still predict academic achievement even after controlling for key socioeconomic variables such as demographics, school attendance, and home educational materials. Meta-analyses testify that non-cognitive constructs such as conscientiousness, self-efficacy, achievement motivation, and test anxiety predict academic achievement and attrition rates over and above the

effects of cognitive ability and socioeconomic status (Poropat, 2009; Robbins, Lauver, Le, Davis, Langley, & Carlstrom, 2004; Seipp, 1991). It seems that students' non-cognitive qualities are important in their own right, and play a vital role in whether students are able to profit from their experience of school.

Given the importance of non-cognitive factors for school success, the goal of the current chapter is to provide a broad overview of the different ways that these non-cognitive constructs can be conceptualized and measured. We begin with a brief explanation of the different types of non-cognitive qualities that have been examined in the literature, focusing primarily on an education context. We then discuss both traditional and innovative measurement methods for indexing these kinds of constructs, and evaluate the strengths and weaknesses of each psychometric approach. Traditional approaches include self- and other-report rating scales and interviews. Novel approaches include situational judgement tests (SJTs), day reconstruction, implicit association tests (IATs), and conditional reasoning tests (CRTs). One commonly stated disadvantage of non-cognitive compared to cognitive assessment is the concern that test-takers are able to distort their responses to create an erroneous impression. We therefore discuss faking in non-cognitive assessment, devoting special attention to whether innovative assessment methodologies can mitigate the effects of faking. In a concluding section, we discuss the existing and potential applications of non-cognitive measurement in educational research, policy, and practice.

What Are Non-cognitive Factors?

"Non-cognitive constructs" is a broad umbrella term in the education literature that refers to a range of student characteristics thought to be distinct from students' intellectual competence and their capacity for mastering the "three Rs" of schoolwork. The extensive research on non-cognitive constructs crosses many different disciplines, including education, educational psychology, social psychology, developmental psychology, personnel psychology, and individual differences (to name just a few). This breadth of research can sometimes result in inconsistent use of construct labels across studies from different disciplines, thus leading to both the "jingle fallacy" (assuming that two constructs are the same because they share the same label) and the "jangle fallacy" (assuming that two constructs differ because they have different labels; e.g., Block, 1995). For

example, the meta-cognition literature uses the term *confidence* to refer to an evaluation of correctness (e.g., "I am confident my answer is right") whereas the positive psychology literature uses the term *confidence* to refer to a positive emotional state (e.g., "I am feeling happy and confident today"). Conversely, the term *integrity* may vary widely in meaning from *intellectual integrity* to through to *absenteeism*. While we acknowledge that terminology is currently not set in stone and may vary slightly from discipline to discipline, we provide a rough taxonomy of some of the most commonly researched non-cognitive constructs, grouped into four domains: (1) attitudes and beliefs, (2) social and emotional qualities, (3) habits and processes, and (4) personality traits.

Attitudes and Beliefs

The first broad group of non-cognitive constructs includes the beliefs that students hold about themselves as learners, the nature of learning, the fairness or supportiveness of the school environment, and their attitudes towards different disciplinary areas and towards school in general. One prominent self-belief system is Dweck and Leggett's (1988) "implicit theories of intelligence." This model proposes that students may hold different types of beliefs about the malleability of intelligence. Entity theorists believe that ability is preset, and cannot be changed through training or practice. In contrast, incremental theorists believe that ability is malleable. As might be expected, entity theorists tend to perform worse at school (Blackwell, Trzesniewski, & Dweck, 2007). This beliefs-and-attitudes grouping of non-cognitive constructs also includes the conscious or unconscious attitudes that students may hold about particular disciplines (e.g., beliefs that mathematics is difficult or that science is fun). Such attitudes may strongly influence students' subsequent behavior and thus their academic success (e.g., Lipnevich, MacCann, Krumm, Burrus, & Roberts, 2011). Students may also hold particular beliefs relating to their self-concept, self-confidence, or self-efficacy, which predict their achievement in various academic and life domains (e.g., Marsh, Byrne, & Shavelson, 1988).

Social and Emotional Qualities

There is a range of non-cognitive constructs that relate to students dealing with their emotions and the emotions of others. Perhaps the most salient and most frequently studied of these constructs is test anxiety (e.g., Sarason, 1984; Zeidner, 1998).

Students who suffer from test anxiety become overwhelmed by the thoughts and physiological sensations of anxiety during assessment situations and thus will be distracted from the task at hand, impairing their performance (e.g., Wine, 1971). Other non-cognitive constructs relating to students' emotions include self-regulation, emotion management, emotional control, coping with stress, and students' emotional states (e.g., MacCann, Fogarty, & Roberts, 2012; MacCann, Wang, Matthews, & Roberts, 2010; Pekrun, Elliot, & Maier, 2006). Research has demonstrated plausible causal pathways that relate greater emotional skills to higher levels of achievement. For example, students with better emotion management tend to use more effective coping strategies in response to academic stressors, which relates to higher levels of academic achievement (MacCann et al., 2011). Social and interpersonal constructs such as teamwork and leadership might also be considered under this broad banner (e.g., Wang, MacCann, Zhuang, Liu, & Roberts, 2009).

Habits and Processes

A further category of non-cognitive constructs relates to particular classes of habits or processes that students engage in when completing academic tasks. For example, students differ in the extent to which they engage in time management practices such as list-making, using time management aids (e.g., a planner or a system of electronic reminders), allocating time to particular tasks, or carefully noting deadlines for assignment due dates. Liu, Rijmen, MacCann and Roberts (2009) demonstrate that time management predicts academic achievement at middle school, and that girls at this age tend to have better time management habits than boys. Similarly, some students may routinely set particular kinds of learning goals, whereas others may not set goals at all, or may set goals that relate to publicly proving their ability rather than learning new things (Grant & Dweck, 2003). Other constructs in this category include organizational skills, study habits, learning strategies, and test-taking strategies (e.g., Crede, & Kuncel, 2008; Liu, 2009). Students' meta-cognitive skills such as self-monitoring and self-evaluation might also be considered part of this broad group of constructs (e.g., Flavell, 1979; Stankov & Lee, 2008).

Personality Traits

Finally, the broad personality domains and narrow personality facets have a long research history

within psychology and education (e.g., Fiske, 1949; Norman, 1963). Personality traits are thought to be relatively stable and long-lasting, and describe an individual's consistent patterns of thoughts, behaviors, and emotions across different situations. Although there are several competing personality models, there is a rough consensus that five broad domains are the best starting points for describing differences between individuals' personalities (e.g., Digman & Inouye, 1986; Tupes & Christal, 1992). These five factors are:

- (1) Extraversion (the tendency to be social, positive, and energetic);
- (2) Agreeableness (the tendency to be kind, truthful and trusting);
- (3) Conscientiousness (the tendency to be detail-oriented, achievement striving, and work-focused);
- (4) Neuroticism (the tendency to experience negative emotions easily, often, and strongly) and
- (5) Openness to Experience (the tendency to be open to new feelings, thoughts, and experiences).

Both conscientiousness and openness to experience show a robust relationship with academic achievement, although only conscientiousness predicts achievement independently of cognitive ability (Poropat, 2009; Trapmann, Hell, Hirn, & Schuler, 2007). In addition to the five broad personality domains, most contemporary models of personality also acknowledge more specific facets of personality that underlie each domain (e.g., Costa & McCrae, 1995; MacCann, Duckworth, & Roberts, 2009). For instance, Costa and McCrae propose six relatively distinct facets of conscientiousness: competence, order, dutifulness, achievement striving, self-discipline, and deliberation. Some research indicates that the narrow facets of personality may be more predictive than the broad domains (e.g., Paunonen & Ashton, 2001).

Each of these different types of non-cognitive factors can be assessed in a variety of ways, and each method of assessment has different strengths and weaknesses. The method of assessment can affect a large variety of test characteristics. These include (but are not limited to):

- (a) the nature of what is measured,
- (b) the potential for response distortion,
- (c) the accuracy of measurement in different groups (e.g., some methods may be more appropriate for young children or test-takers with developing English-language literacy);

- (d) the type and variety of feedback that can be given;
- (e) the feasibility of large-scale testing and group assessment, compared to individual assessment;
- (f) the cost of testing in terms of time, money, required equipment, assessor-training, and assessment development expenses;
- (g) the reliability of the tests;
- (h) the ease of building interventions or development plans targeting the measured construct;
- (i) the sensitivity of the measure to changes over time;
- (j) the potential for adverse impact, particularly if the assessment is used for high-stakes purposes such as selection; and
- (k) test-taker reactions and engagement with the testing process.

In the paragraphs below, we describe and evaluate some of the most commonly used assessment methodologies, as well as some of the innovative and emerging methods that may be used in future assessments.

Traditional Non-cognitive Assessment Techniques

Self-Assessments

Self-assessments are undoubtedly the most widely used approach for gauging students' non-cognitive characteristics. These uses include: evaluating the effects of training; program evaluation; outcomes assessment; research; and large-scale, group-level national and international comparisons, to name a few. Indeed, most insights concerning the relationship between non-cognitive qualities and educational (or for that matter, work-related) outcomes come from research, practice, and policy conducted with self-report questionnaires.

GENERAL APPROACH

Self-assessments usually ask individuals to describe themselves by answering a series of standardized questions. The answer format is generally a Likert-type rating scale, but other formats may also be used (such as Yes/No or constructed response). Typically, questions assessing the same construct are aggregated; this aggregate score serves as an indicator of the relevant non-cognitive domain. The variety of constructs that can plausibly be assessed with self-reports are myriad. At a broad or abstract level these include personality, values, beliefs, and affect. Examples of specific

constructs that may be assessed with self-reports include communication skills, time management, teamwork, leadership, self-regulation, self-efficacy, and altruism. (Table 33.1 in this chapter includes a selection of sample items indicative of this approach.)

Self-assessments are a relatively pragmatic, cost-effective, and efficient way of gathering information about the individual. However, many issues must be taken into account when one is developing a psychometrically sound questionnaire, and there is a large literature on a wide variety of such topics. The optimal number of points on a scale, scale point labels, the inclusion of a neutral point, alternative ordering, and other test characteristics have been widely analyzed and examined in the literature (e.g., Krosnick, Judd, & Wittenbrink, 2005). For instance, studies reveal that response scale format influences individuals' responses (Rammstedt & Krebs, 2007), while the inclusion of negatively keyed questions (to avoid acquiescence) is considered controversial, especially with younger children (e.g., Barnette, 2000; DiStefano & Motl, 2006). Respondents vary in their use of the scale—for example, young males tend to use extreme answer categories (Austin, Deary, & Egan, 2006), as do Hispanics (Marin, Gamba, & Marin, 1992); and in general, there are large cultural effects in response style (Harzing, 2006; Lipnevich et al., 2011).

Respondents can fake their responses on self-assessments to appear more attractive to a prospective educational institution or employer, or to avoid remedial programs (e.g., Griffith, Chmielowski & Yoshita, 2007; Viswesvaran & Ones, 1999; Zickar, Gibby & Robie, 2004). Researchers have identified several promising methods for collecting data through self-reports while reducing fakeability. These include giving real-time warnings (Sackett, 2006), using a forced-choice format (Stark, Chernyshenko, & Drasgow, 2005), and using one's estimates of how others will respond to help control for faking (Prelec, 2004; Prelec & Weaver, 2006). However, evidence for the effectiveness of these procedures in controlling for faking remains to be demonstrated unequivocally (Converse, Oswald, Imus, Hedricks, Roy, & Butera, 2008; Heggstad, Morrison, Reeve, & McCloy, 2006). We consider the issue of fakeability in greater detail in a later section of this chapter.

THEORY OF PLANNED BEHAVIOR (TPB) AS A FRAMEWORK FOR DEVELOPING SELF-ASSESSMENTS

Self-assessments are based on a number of different conceptual frameworks, including those based

Table 33.1 Self-Report: Examples of Constructs, Items, and Response Scales

Construct	Sample Items	Response Scale
1 Achievement Striving	1. I detect mistakes. 2. I do just enough work to get by (R).	Five-point Likert scale: "Not At All Like Me" (1) to "Very Much Like Me" (5).
2 Conscientiousness	1. I am always prepared. 2. I work hard.	Five-point Likert Scale: "Very Inaccurate of Me" (1) to "Accurate" (5).
3 Goals	1. I focus on the happy ending. 2. I fully focus on the obstacles (R).	Seven-point Likert scale: "Never" (1) to "Always" (7).
4 Grit	1. I am diligent 2. Failures double my motivation to succeed.	Five-point Likert scale: "Not at all like me" (1) to "Very much like me" (5).
5 Learning Strategies	1. When I study, I try to figure out which parts of the material I need to study most. 2. When I do my homework, I check to see whether I understand the material.	A four-point Likert-type scale: Never or Hardly Ever (1), Sometimes (2), Often (3), Always or Almost Always (4).
6 Self-Efficacy	1. I am certain that I can accomplish my goals. 2. I can handle whatever comes my way.	A four-point Likert-type scale: Hardly Ever (1), Sometimes (2), Often (3), or Almost Always (4).
7 Academic Motivation	1. I do only as much work as I have to for the grade I want. 2. I put little effort into my classes (R).	A four-point Likert-type scale: Strongly Agree (4), Agree (3), Disagree (2), Strongly Disagree (1).
8 Feelings About School Life	1. When doing after-school activities, I have felt nervous (R). 2. When doing homework, I have felt confident.	A four-point Likert-type scale: Never or Rarely (1), Sometimes (2), Often (3), Usually or Always (4).
9 Anxiety	1. I worry about things. 2. Fear for the worst.	Yes/No
10 Leadership	1. Can talk others into doing things. 2. Wait for others to lead the way.	Five-point Likert scale: "Not at all like me" (1) to "Very much like me" (5).

Notes: (R) refers to items that are reverse-keyed for respective scales of the instruments.

on clinical criteria, lexical analysis, and psychological theory. TpB has been particularly effective in serving as a framework for the development of assessments of individuals' attitudes (see Table 33.2 in this chapter for examples of items measuring TpB components). Only recently has the theory of planned behavior been applied in educational contexts (see Davis, Ajzen, Saunders, & Williams, 2002; Lipnevich et al., 2011). The initial findings appear promising. Hence, a brief overview follows.

The TpB is based on the psychological theory of reasoned behavior (Ajzen, 1991, 2002). The TpB posits that the central determinant of volitional behavior is one's intention to engage in that behavior. The theoretical model of the TpB is shown in Figure 33.1. Ajzen (1991) further proposes three independent determinants of behavior that exert

their effects through intentions. These are: (1) attitudes, (2) subjective norms, and (3) perceived behavioral control. *Attitudes* are defined as the overall positive or negative evaluation of the behavior. In general, the more favorable the attitude towards the behavior, the stronger the individual's intention is to perform it. Subjective norms assess the social pressures on the individual to perform or not to perform a particular behavior. Finally, perceived behavioral control provides information about the potential constraints on action as perceived by the individual (Armitage & Conner, 2001).

Several meta-analyses and literature reviews show support for the general principles underlying the TpB model (see Ajzen, 1991; Armitage & Conner, 2001). Studies reveal that the TpB accounts for 27 percent and 39 percent of the variance in behavior

Table 33.2 Self-Report: Theory of Planned Behavior–Based Assessment of Attitudes Toward Mathematics (after Lipnevich et al., 2011).

TpB Component	Sample Item	Response Scale
1. Attitudes	I enjoy studying math.	(1) Strongly Disagree
2. Subjective Norm	My friends think that math is an important subject.	(2) Disagree (3) Neither Agree nor Disagree
3. Perceived Behavioral Control	If I invest enough effort, I can succeed in math.	(4) Agree (5) Strongly Agree
4. Intentions	I will try to work hard to make sure I learn math.	

and intention, respectively (Armitage & Conner, 2001; Sheeran, 2002). Davis et al. (2002) used the TpB to successfully predict high school completion (Davis, Ajzen, Saunders, & Williams, 2002). Results revealed that the TpB questionnaire significantly predicted high school completion. The model was a good fit to the data, and attitudes, subjective norms, and perceived control accounted for 51 percent of the variance in the intention to complete the present school year. Furthermore, attitudes, subjective norms, and perceived control all significantly predicted intention.

Another study that applied TpB to an educational context was conducted by Lipnevich et al. (2011). The researchers examined the effectiveness of the TpB in predicting students' mathematics performance. They found that between 25 percent and 32 percent of the variance in mathematics grades could be explained by TpB components. Moreover, 17 percent of the variation in test grades can be

explained by the TpB over and above the effects of mathematics ability test scores. So, development and implementation of TpB-based self-assessments may be instrumental in predicting a number of meaningful educational outcomes. Additionally, researchers suggest that the TpB has tremendous potential to inform the development of behavior-change interventions (see e.g. Armitage & Conner, 2001; Hardeman, Johnston, Johnston, Bonetti, Wareham, & Kinmonth, 2002; Rutter, 2000).

Other-Ratings

Other-ratings are assessments in which others (e.g., supervisors, trainers, colleagues, friends, faculty advisors, coaches, etc.) rate individuals on various non-cognitive skills. This method has a long history, and numerous studies have been conducted that employed this methodology to gather information (e.g., Tupes & Christal, 1961/1992). Other-ratings have an advantage over self-assessments in that they

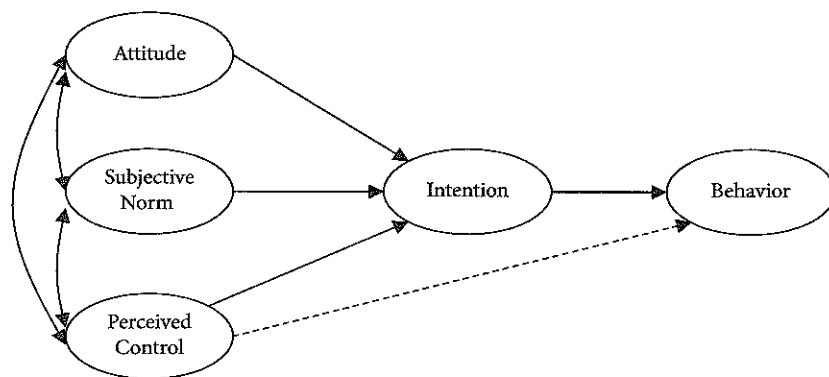


Figure 33.1 Representation of the Components of the TpB model.

Key: *Attitudes*—the overall evaluation of whether a behavior is positive or negative (based on prior behavioral contingencies). *Subjective Norms*—the perceived social pressure to perform the behavior. *Perceived Control*—the person's estimate of his or her capacity to perform the behavior. *Intentions*—the readiness or willingness to perform the behavior.

preclude socially desirable responding, although they are prone to rating biases. Self- and other-ratings do not always converge (Oltmanns & Terkheimer, 2006), but other-ratings have been demonstrated to often be more effective in predicting a range of educational outcomes, compared to self-ratings (Kenny, 1994; Wagerman & Funder, 2006).

TEACHER- AND PARENT-RATINGS

Teacher- or parent-ratings of personality have been widely used for gauging younger students' non-cognitive characteristics due to concerns that children lack the cognitive ability and/or psychology-mindedness to self-rate on personality instruments (e.g., Hendriks, Kuyper, Offringa, & VanderWerf, 2008). Other-ratings are also most appropriate when used to evaluate individuals with low verbal ability. For instance, the Child Behavior Check List (CBCL; Achenbach, 1991)—completed by teachers and/or parents—has been successfully employed to assess behavioral and emotional competencies (e.g., Anxious/Depressed, Rule-breaking Behavior) of children aged 6 to 18 years. Similarly, the Parent Rating Scale of the Behavioral Assessment System for Children (currently in its second edition—BASC-2) has been widely used to gauge adaptive and behavior problems of children between the ages of two and 21 (Reynolds & Kamphaus, 2004). Both CBCL and BASC-2 has been shown to capture a range of problems and competencies that would be difficult—if not impossible—to capture through self-reports.

MacCann, Lipnevich, and Roberts (2013) conducted a series of studies to compare parent judgments of personality with sixth to eighth-grade students' self-assessments. The results from three studies suggested that parent-ratings were both more reliable and more useful in predicting academic achievement than students' self-reports. This large difference in utility between self- and parent-reports of personality is noteworthy: Parent-reported Conscientiousness explained over twice as much variation in grades as students' self-reported Conscientiousness. Although such results might be used to justify the idea that younger children (in their preteens and early teens) might lack the psychology-mindedness or cognitive ability to accurately self-rate on personality questionnaires, it is worth comparing current results to studies of self- versus other-reports in older teenagers and adults. Thus, research studies demonstrate that peer-, co-worker-, supervisor-, and customer-ratings may be more reliable and more predictive of

valued outcomes than self-ratings, particularly for Conscientiousness (e.g., Mount, Barrick, & Strauss, 1994; Small & Diefendorff, 2006).

There are several possible reasons that parent-reports might differ from self-reports. John and Robins (1993) speculate that self-reporting differs from other-reporting in three main ways: (1) ego involvement is implicated in self-reports but not other-reports; (2) perspectives are different (the other-report is from an external observer, whereas the self is actively involved in the phenomena that personality items ask about); and (3) the self has access to privileged information such as previous experiences, internal thoughts, values, and intentions that are not available to others. If point (1) is responsible for the greater prediction by parent-reports, response bias in self-reports might attenuate correlations with grades. However, large-scale results from personnel psychology demonstrate that response bias does not actually affect the predictive power of personality tests (Barrick, & Mount, 1996; Ones, Viswesvaran, & Reiss, 1996). Moreover, when comparing mean differences in self- versus parent-reports, MacCann et al. (2013) demonstrated the opposite effect—parent-ratings resulted in higher means (i.e., were more flattering) than self-ratings. MacCann et al.'s study showed that score inflation in parent- versus child-reports might relate to different processes, with parent-reports prone to self-deceptive denial, whereas inflation of self-reports appears due to self-deceptive enhancement. Newspapers are full of illustrative examples of parents who deny that their children could possibly have behaved in such an immoral or unrestrained fashion as to commit crimes, for example, despite all the evidence to the contrary. When the other-rater shares a close emotional bond with the individual being assessed, the general rule that score inflation will be greater for self- than other-reports appears controvertible.

Point (3) might also account for differences in parent and self-report prediction. Given that parents do not have access to privileged internal information, they may need to use cues from the child's academic outcomes (which they know) to answer questions about the child's personality. That is, parents may make estimates on available information (which may be drawn from other constructs) when answering questionnaire items that refer to this privileged or internalized information. The implication of this finding for test development is that parent-reports (and other-reports more generally) should refer to observable information or facts to avoid the risk of criterion contamination.

However, parent-reports do have the advantage over self-reports in terms of point (2). A more external perspective might be viewed as a more objective reporting of the facts and so might result in more accurate measurement of the construct. Overall, parent- and teacher- reports are indispensable when assessing non-cognitive characteristics of younger students or students with limited verbal abilities.

LETTERS OF RECOMMENDATION

Letters of recommendation represent a specific form of other-rating and have been extensively used in a broad range of educational (e.g., Vannelli, Kuncel, & Ones, 2007) and workplace contexts (e.g., Arvey, 1979). Letters of recommendation provide stakeholders with detailed information about applicants' past performance, with the writers' opinion about the applicant being presented in the form of an essay. One major drawback of letters of recommendation is that they are not in a standardized format: different letter-writers may include or exclude qualitatively different types of information, so it is difficult to judge one letter against another. Walters, Kyllonen, and Plante (2003, 2006) developed a standardized format for such letters to counter this perceived problem (see Table 33.3 in this chapter for sample items). Initially termed the Standardized Letter of Recommendation, and now the Educational Testing Service (ETS)* Personal Potential Index (ETS, 2009), this assessment system prompts faculty members to respond to specific items using a Likert scale, in addition to eliciting their open-ended comments. It has been used operationally at ETS for selecting summer interns and fellows (Kyllonen & Kim, 2004; Kim & Kyllonen, 2008), through Project 1000 for the selection of graduate student applicants (Liu, Minsky, Ling, & Kyllonen, 2007), and since 2009, it has supplemented the Graduate Record Examination (GRE; see ETS, 2009; Kyllonen, 2008). Several research teams are currently collecting and analyzing data to address questions of validity and predictive power of letters of recommendation, but preliminary results provide some evidence for the reliability and validity of this method of measurement.

Biodata

Biographical data (biodata) have been explored for college admissions in the United States (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004), Chile (Delgalarrando, 2008), and other countries. Biodata has also been a standard methodology for assessing constructs such as opportunity to learn

Table 33.3 The ETS* Personal Potential Index: Sample Items

	Construct and Sample Items	Response Scale
1	Knowledge and Creativity Has a broad perspective on the field Produces novel ideas	(1) Below average (2) Average (3) Above average
2	Ethics and Integrity Is among the most honest persons I know Maintains high ethical standards	(4) Outstanding (5) Truly exceptional (7) Insufficient opportunity to evaluate
3	Planning and Organization Sets realistic goals Meets deadlines	
4	Resilience Accepts feedback without getting defensive Works well under stress	
5	Teamwork Supports the efforts of others Works well in group settings	
6	Communication Skills Speaks in a clear, organized, and logical manner Writes with precision and style	

and socioeconomic status in large-scale national and international group-level comparative studies (e.g., the National Assessment of Educational Progress, Programme for International Student Assessment, and Trends in International Mathematics and Science Study) (Chapter 6, this volume). Biodata are typically obtained by asking standardized questions about individuals' past behaviors, activities, or experiences. Respondents are typically offered multiple-choice answer options or are requested to answer questions in an open format (e.g., "state frequency of behavior"). Baird and Knapp (1981; see also Stricker, Rock, & Bennett, 2001) developed a biodata (documented accomplishments) measure that produced scores for six scales: Academic Achievement, Leadership, Practical Language, Aesthetic Expression, Science, and Mechanical.

Jackson, Wood, Bogg, Walton, Harms, & Roberts (2010) demonstrated that biodata approach can be effective when assessing individuals' personality. The researchers attempted to identify the behavioral component of conscientiousness and to specify a

relatively large pool of behaviors that represent this personality facet. They developed and validated the Behavioral Indicators of Conscientiousness (BIC) and showed that the lower-order structure of conscientious behaviors (as assessed by BIC) is nearly identical to the lower-order structure obtained from extant self-report measures. Furthermore, the researchers used a daily-diary method to validate the BIC against frequency counts of conscientious behavior and found that behaviors assessed with BIC were strongly related to behaviors assessed daily through a diary method. The findings of Jackson et al. (2010) allow for a conclusion that may be extended to the biodata method in general: "Reports of past behavior are at least partially valid, mitigating a criticism often applied to self-reports of behavior" (p. 7).

Measures of biodata show incremental validity beyond SAT scores and the Big Five personality scores in predicting students' performance in college (Oswald, et al., 2004). Biodata may offer a less fakeable method of assessment than standard self-report scales, as there are several test characteristics that can be implemented to minimize faking (e.g., Dwight & Donovan, 2003; Schmitt, Oswald, Kim, Gillespie, & Ramsay, 2003). These include asking students to elaborate on the biodata details (e.g., "What was the name of the last foreign movie you saw?") or triangulating results obtained with alternative measurement approaches (e.g., other-reports). A sample biodata item is presented in Table 33.4 in this chapter.

Interviews

Interviews are the most frequently used method of personnel-selection in industry (Ryan, McFarland, Baron, & Page, 1999) and in clinical practice (Meyer et al., 2001), but they are also used for admissions, promotions, scholarships, and other awards in educational contexts (Goho & Blackman, 2006; Hell, Trapmann, Weigand, & Schuler, 2007). Interviews vary in their content and structure. In a

structured interview, questions are prepared before the interview starts. An unstructured interview simply is a free conversation between an interviewer and interviewee giving the interviewer the freedom to adaptively or intuitively switch topics. Research has shown that unstructured interviews lack predictive validity (Arvey & Campion, 1982) or show lower predictive validity than structured interviews (Schmidt & Hunter, 1998).

Structured interviews can be divided into three types: the behavioral description interview (BDI; Janz, Hellervik, & Gilmore, 1986), situational interview (SI; Latham, Saari, Pursell & Campion, 1980), and multi-modal interview (MMI; Schuler, 2002). The behavioral description interview involves questions that refer to the candidate's past behavior in real situations. The situational interview uses questions that require that interviewees imagine hypothetical situations (derived from critical incidents) and state how they would act in such situations. The multi-modal interview combines the two approaches and adds unstructured parts to ensure high respondent acceptance.

Meta-analyses of the predictive validity of interviews for job performance (Huffcutt, Conway, Roth, & Klehe, 2004; Marchese & Muchinski, 1993; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Schmidt & Hunter, 1998) show that structured interviews: (a) are good predictors of job performance (corrected correlation coefficients range from 0.45 to 0.55); (b) they add incremental validity above and beyond general mental ability; and that (c) behavior description interviews show a higher validity than situational interviews.

Similarly, in educational contexts, interviews have been deemed moderately effective for the prediction of meaningful outcomes. For example, Goho and Blackman (2006) investigated the effectiveness of using selection interviews for admissions into medical schools, conducting meta-analyses to predict academic achievement and clinical performance. The mean effect size for studies examining the predictive power of interviews for academic success was 0.06 (95% confidence intervals 0.03–0.08), indicating a very small effect, whereas the sample of studies for predicting clinical success had a mean effect size of 0.17 (95% confidence intervals 0.11–0.22), indicating modest positive predictive power.

Wilkinson et al. (2008) report similar findings. Their study examined how well prior academic performance, admission tests, and interviews predicted academic performance in a graduate medical school. The researchers found that medical school grade

Table 33.4 Sample Biodata Item (from the Leadership scale of Oswald et al., 2004, p. 204).

Item	Response
How many times in the past year have you tried to get someone to join an activity in which you were involved or leading?	(A) never (B) once (C) twice (D) three or four times (E) five times or more

point average (GPA) was most strongly correlated with prior academic performance (e.g., for overall score, partial $r = 0.47$; $p < 0.001$), followed by interviews (partial $r = 0.12$). Interestingly, whereas the relationship between GPA and performance weakened from Year 1 to Year 4, the association between interview score and performance increased from Year 1 to Year 4. Considering that the admissions interviews mostly focus on assessing candidates' non-cognitive characteristics (i.e., their motivation to become doctors, interest, drive, etc.), these findings attest to the effectiveness of interviews as methods for gauging students' non-cognitive skills. Thus, interviews were found to be better predictors of both medical school GPA and clinical practice than the Graduate Australian Medical School Admissions Test (Wilkinson et al., 2008), a finding reminiscent of the Swedish enlistment study conducted by Lindqvist and Vestman (2011).

Innovative Non-cognitive Assessment Techniques

Situational Judgement Tests (SJTs)

SJTs are a type of test where individuals are presented with a situation and then select either the most appropriate response or their typical response out of a list of possible choices (see Table 33.5, this chapter, for sample items). SJTs can be text-based or presented through multimedia, and responses can be multiple-choice (i.e., pick the best response), constructed response (i.e., provide a response to this situation), or ratings (i.e., rate each response for its effectiveness, on a Likert-type scale) (see, e.g., McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). SJTs represent fairly simple, economical simulations of relevant academic- (or job-) related tasks (Kyllonen & Lee, 2005).

SELF-RATED SJTS

SJTs have several advantages over traditional self-assessment instruments. First, SJTs may be developed to reflect more subtle and complex judgement processes than are possible with conventional tests. Carefully constructed, the methodology of the SJT enables the measurement of many relevant attributes of applicants, including social competence, communication skills, critical thinking, and leadership, to name a few (e.g., Oswald et al., 2004; Waugh & Russell, 2003). By getting at these hard-to-measure constructs, SJTs carry the possibility of overcoming the validity ceiling found for conventional cognitive assessments in personnel selection and college admissions. Second, SJTs appear to be associated

Table 33.5 Situational Judgement Test Item: Teamwork Assessment

Item stem	You are part of a study group that has been assigned a large presentation for class. As you are all dividing up the workload, it becomes clear that both you and another member of the group are interested in researching the same aspect of the topic. Your colleague already has a great deal of experience in this area, but you have been extremely excited about working on this part of the project for several months. Which of the following is the best approach to dealing with this situation?
Responses	<p>(A) Flip a coin to determine who gets to work on that particular aspect of the project.</p> <p>(B) Insist that, for the good of the group, you should work on that aspect of the project because your interest in the area means you will do a particularly good job.</p> <p>(C) Compromise your preferences for the good of the group and allow the other person to work on that aspect of the project.*</p> <p>(D) Choose a different group member to work on that aspect of the project so that no one person is privileged over another.</p>

with less adverse impact on (ethnic) minorities. Of relevance in this context, reduced subgroup differences have indeed been found with SJTs (e.g., McDaniel et al., 2001). Third, SJTs can be used in training sessions to provide a student or prospective employee with feedback on his or her competencies in the domain of interest. Finally, SJTs appear to be less susceptible to faking than are self-assessments, where the improvement due to incentives can be up to a full standard deviation.

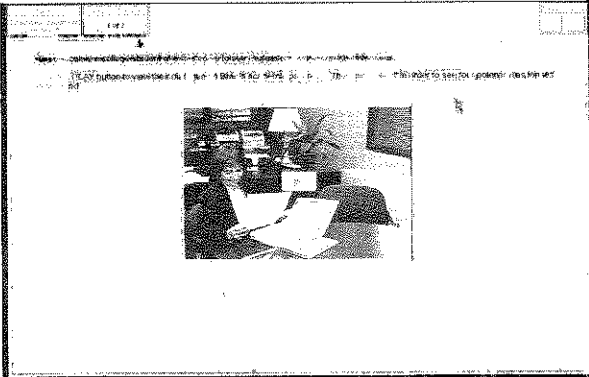
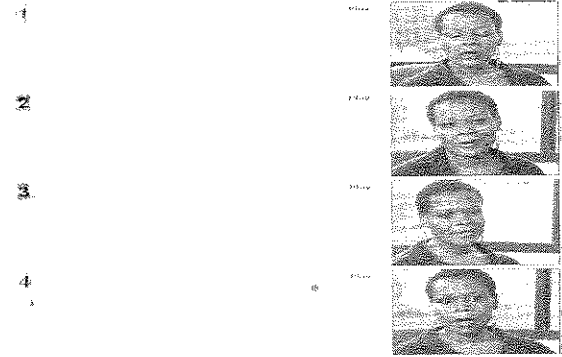
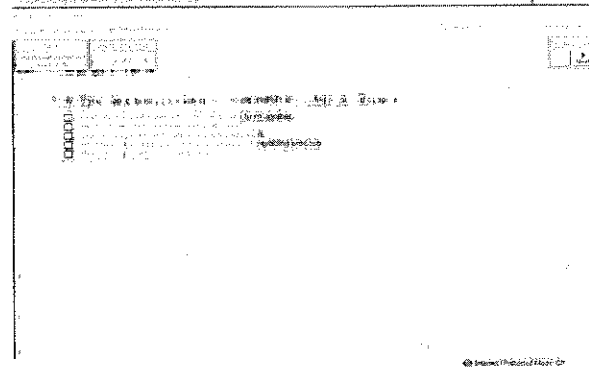
SJTs have been shown to predict a range of important outcomes such as college success (Lievens & Coetsier, 2002; Oswald et al., 2004) and leadership (Krokos, Meade, Cantwell, Pond, & Wilson, 2004). Although applications in education have been relatively limited, using SJTs in educational domains is certainly on the rise (Lievens, Buyse, & Sackett, 2005; MacCann et al., 2010; Oswald et al., 2004; Wang et al., 2009). This trend is partly due to the fact that there is more and more evidence that SJTs have high construct validity, both of predictive and consequential

nature (e.g., Erienne & Julian, 2005; Sternberg et al., 2000). A recent study of high school students compared showed that an SJT of teamwork showed a higher correlation with GPA than a self-report rating scale of teamwork (Wang et al., 2009).

Depending on the information delivery mode, SJTs are typically dichotomized into text-based and multimedia-based types. Text-based SJTs fall into the category of traditional SJTs in the sense that they are presented in a paper-and-pencil format where scenarios and response options are in written form. Quite differently, multimedia SJTs use multimedia

technology to present scenarios and sometimes also response options in video format (Lievens et al., 2005; McHenry & Schmitt, 1994; Olson-Buchanan & Drasgow, 2006; Weekley & Jones, 1997). Recent meta-analytic results demonstrate that multimedia SJTs show stronger criterion-related validity than written SJTs for predicting interpersonal skills (Christian, Edwards, & Bradley, 2010), and appear most effective when used to assess students' affective characteristics. A figural representation of a multimedia SJT item is presented in Table 33.6 in this chapter.

Table 33.6 Figural Representation of a Typical Multimedia SJTEA Item

Scenario	
Response scale	
Response justification	

There are several reasons for the above-mentioned finding. First, multimedia SJTs enable the assessment of certain emotional abilities that cannot easily be assessed with other methodologies that are limited by the delivery medium. For instance, written SJTs can only provide verbal content, but multimedia SJTs represent a richer medium because they can present many more social cues, including verbal, nonverbal, and paralinguistic information.

Secondly, the use of multimedia technology enhances the degree to which the test mirrors the real environment, also referred to as *stimulus fidelity* (i.e., the extent to which the assessment task and context mirror those actually present in real life, Callinan & Robertson, 2000). Higher stimulus fidelity of multimedia SJTs is related to (a) enhanced ecological validity of the test (Chan & Schmitt, 1997), (b) more favorable test-taker reactions (Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000), and (c) better prediction of meaningful outcome variables (Christian et al., 2010; Lievens & Sackett, 2006).

Thirdly, the use of multimedia technology does not rely excessively on verbal ability, a problem that has characterized the field to date. For example, written SJTs require the understanding and interpretation of text, making them dependent on text comprehension. The fact that text-based SJTs assume fairly advanced levels of reading comprehension may constitute a source of construct-irrelevant variance and may lead to higher correlations with cognitive ability. Contrary to text-based SJTs, multimedia SJTs do not require as much reading comprehension, instead more clearly targeting focal processes such as perceiving emotions and understanding the causes of these emotions. Indeed, recent studies confirm that multimedia SJTs have lower correlations with cognitive ability than do text-based SJTs (Lievens & Sackett, 2006).

In sum, SJTs represent a promising method for assessment of students' non-cognitive characteristics, with multimedia SJTs offering something that text-based assessments do not. Despite relatively high cost associated with their production (e.g., hiring actors and videotaping) and special requirements for administration (e.g., computers), we predict that multimedia SJTs will become an increasingly popular tool for measuring students' affective characteristics in a range of educational contexts (e.g., as supplements to college or graduate school admission interviews). An example of an SJT item is presented in Table 33.5.

OTHER-RATED SJTs

It is possible to administer SJTs in other-report format. That is, an observer such as a parent or teacher would be presented with a particular situation, and would judge what the target student would do in that situation. An example of the other-rated SJT item is presented in Table 33.7. MacCann, Wang, Matthews, and Roberts (2010) demonstrated that SJTs can be reliably administered in other-report format. This study used both self-report and parent-report SJTs to assess middle schools students' emotion-management skills. They found that although self- and parent-judgements were only weakly related, they both independently predicted valued criteria such as school grades, life satisfaction, and emotional reactions (both positive and negative) to the school environment. These findings support the idea that the non-shared variation between self- and parent-judgements is not random error, but represents different aspects of the phenomenon under investigation (in this case, emotion management). The researchers suggest that self- and parent-evaluations may index the frequency of different types of emotion management strategies. Parent-evaluations might relate to strategies involving interaction with others (e.g., seeking social support, talking through the issues), whereas self-evaluations might relate to strategies involving kind and sympathetic feelings towards others (e.g., expressing sympathy to diffuse tension, complimenting others). Other-rated SJTs represent a very promising approach for assessing non-cognitive characteristics, as they combine the benefits of SJTs

Table 33.7 Situational Judgement Test Item: Parent-Report of Teamwork Assessment (after MacCann et al., 2010).

Item stem	Your child and a classmate, James, sometimes help each other with homework. After your child helps James on a difficult project, the teacher is very critical of this work. James blames your child for his bad grade. Your child responds that James should be grateful, because helping him was a favor. What would your child do in this situation?
Responses	(A) Tell James that from now on he has to do his own homework. (B) Apologize to James. (C) Tell James "I am happy to help, but you are responsible for what you turn in." (D) Don't talk to James.

(e.g., high ecological validity of situations) and other-reports (e.g., suitable for younger kids and those with limited verbal abilities).

Time Use: Day-Reconstruction Method

A relatively new behavioral science domain concerns how people use their time. An assessment technique is the Day-Reconstruction Method (DRM; Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). The DRM assesses how people spend their time and how they experience the various activities and settings of their lives. It combines features of two other time-use techniques, time-budget measurement (the respondent estimates how much time is spent on various categories of activities) and experience sampling (the respondent records his or her current activities when prompted, to do so at random intervals throughout the day). The DRM requires that participants systematically reconstruct their activities and experiences of the preceding day with procedures designed to reduce recall biases.

When using the DRM, a respondent first recreates the previous day by producing a confidential diary of events. Confidentiality encourages respondents to include details they may not want to share through any other assessment approach (such as an interview). Next, respondents receive a standardized response form and use their confidential diary to answer a series of questions about each event, including (1) when the event began and ended, (2) what they were doing, (3) where they were, (4) whom they were interacting with, and (5) how they felt on multiple affect dimensions. The response form is returned to the researcher for analysis. In addition, respondents answer a number of demographic questions.

Respondents complete the diary before they are informed about the content of the standardized response form, so as to minimize biases. A study of 909 employed women showed that the DRM closely corresponds with experience sampling methods (Kahneman et al., 2004). The DRM is a time-consuming and intrusive form of assessment that requires a significant effort from respondents. More research is needed to capture psychometric qualities of the method. However, initial evidence suggests that this method is effective in assessing characteristics otherwise difficult to capture (Belli, 1998; Kahneman et al., 2004). Moreover, the method appears generalizable to high school and college populations. For example, Roberts et al. (2011) report similar findings with the DRM for 131 college freshman as found for employee

samples. In particular, participants reported significantly greater positive affect while engaging in hobbies or socializing than attending class or completing homework. In this study, the DRM was also substantially correlated (i.e., r 's exceeding 0.40 for each activity) with a self-assessment of psychological well-being and a situational judgement test of emotional management.

Writing Samples

Chung and Pennebaker's (2007) analysis of writing samples as a gateway to personality is based on the idea that what we write and say, as well as how we write and say it, reflects our personality. This stream of research involves correlating words and word types from open-ended writing (e.g., emails) with personality and behavioral measures. Research suggests that the use of particular function words (e.g., pronouns, adjectives, articles) is related to individuals' affective states, reactions to stressful life events, social stressors, demographic factors, and biological conditions (Chung & Pennebaker, 2007; Mehl & Pennebaker, 2003). For example, the use of "I" is associated with depression, and speaking to a superior, based on email correspondence. Moreover, word choices can be used to detect deception (Hancock, Curry, Goorha, & Woodworth, 2004; Newman, Pennebaker, Berry, & Richards, 2003).

Vast volumes of materials are available to explore this research program further (especially given the preponderance of social networks and the archival capabilities that are part of the Internet), while the availability of inexpensive automated classification tools provides noteworthy research opportunities to continue to identify relationships between written communication and non-cognitive skills. The magnitude of correlations found tends to be quite low, but the method's low cost and unobtrusiveness suggests that it may lead to future applications in psychological testing. Although most of this research has focused on adult behaviors, the amount of writing that needs to be completed in secondary and tertiary education poses some intriguing possibilities for the assessment of non-cognitive skills during childhood and adolescence.

The Thorny Issue of Response Distortion

A potential problem with using non-cognitive assessments for high-stakes purposes such as educational selection is that test-takers may try to "fake" high scores in order to get into the course. Even an exceptionally lazy person is unlikely to agree with the statement "I am lazy" if they are answering

this item as part of a college admissions process. Meta-analytic research from personnel psychology supports this intuitive idea with hard empirical data. Viswesvaran and Ones' (1999) meta-analysis demonstrates that people can fake personality tests when asked to do so, raising their personality test scores by the equivalent of 7 to 14 IQ points. Furthermore, research suggests that between 20 percent and 40 percent of people actually do fake when taking personality tests for selection purposes (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006).

Psychologists have long realized that people are prone to exaggerate on rating-scale tests in order to get a better score and so get into the school of their choice or procure a better job. The predominant approach to dealing with faking has been to detect fakers with lie scales, also known as *response distortion scales* (e.g., Crowne & Marlowe, 1960; Paulhus, 1998). Lie scales intermix items such as "I never swear," or "I always pick up my litter" with focal personality items. Fakers are identified by their high scores on lie scales. Although lie scales are immensely popular and commonly used, there is a growing consensus that they do not work (Dilchert, Ones, Viswesvaran, & Deller, 2006; Ellingson, Sackett, & Hough, 1999; MacCann, Zeigler, & Roberts, 2011). Both logic and empirical evidence suggest that people scoring high on lie scales might actually be exemplary individuals who always pick up their litter and never swear. Correlations of lie scores with substantive personality traits like conscientiousness and agreeableness suggest that it is not liars, but nice, kind, hard-working people who are caught out by lie scales (e.g., Li & Bagger, 2006).

Given that lie scales do not seem to catch the liars, there are two broad strategies for dealing with response distortion when assessing non-cognitive constructs. First, non-cognitive constructs, particularly as measured with self-reported rating scales, simply cannot be used as selection criteria for high-stakes purposes. For example, it may be reasonable for medical schools to exclude applicants with very low scores on empathy, given that these applicants might make poor healthcare professionals. However, it would be unreasonable to select only the top applicants based on empathy, as the very top scores on a rating-scale measure of empathy might be fakers who displace genuinely empathetic applicants. That is, using non-cognitive scores to screen out wildly inappropriate applicants is still a useful exercise, even if some people fake.

The second strategy to combat faking in non-cognitive assessments is to use a range of

innovative assessment methods that show greater resistance to faking than standard rating scales. Implicit measurement techniques are chief among these new methods. In implicit measurement, the measurement objective is not obvious to the participants, such that these measures may be less susceptible to faking (e.g. Ziegler, Schmidt-Atzert, Buehner, & Krumm, 2008). Implicit measurement techniques include the implicit association test (IAT) paradigm, as well as the conditional reasoning test (CRT) paradigm, which we describe below. Other testing methods that may reduce faking include forced-choice assessment and the Bayesian truth serum. Many of these measurement techniques are still in their infancy, with limited empirical data to provide evidence of their validity and their non-fakeability. In the paragraphs below, we describe each of these techniques in turn and outline the empirical evidence for these techniques as viable methods to measure non-cognitive constructs.

Implicit Assessments: Implicit Association Tests (IATs)

The implicit association test (Greenwald, McGhee, & Schwartz, 1998) has become an incredibly popular method for researching non-cognitive factors, particularly attitudes, having been examined in many hundreds of empirical studies (see Greenwald, Nosek, & Sriram, 2006). IATs record the reaction time it takes to classify stimulus pairs (e.g., word, picture), which is then treated as an indirect measure of whether a participant sees the stimuli as naturally associated. IATs thus measure the strength of implicit associations to gauge attitudes, stereotypes, self-concepts, and self-esteem (Greenwald, Banaji, Rudman, Farnham, Nosek, & Mellor, 2002; Greenwald & Farnham, 2000).

IATs generally exhibit reasonably good psychometric properties. Meta-analyses have revealed high internal consistencies (0.80 to 0.90) (Hofmann et al., 2005), although somewhat lower test-retest reliabilities (0.50 and 0.70), which is a common finding in reaction time research. IATs predict a wide variety of criteria, particularly spontaneous (as opposed to controlled) behavior (Bosson, Swann, & Pennebaker, 2000; Gawronski & Bodenhausen, 2006; McConnell & Leibold, 2001). However, to our knowledge, they have not been used in studies of educational outcomes, although there is an emerging literature using this method to explore the attitudes of children and adolescents to health-related behaviors (e.g., smoking; see Andrews, Hampson, Greenwald, Gordon, &

Widdop, 2010). The Hofmann et al. meta-analysis estimated the correlation between implicit (IATs) and explicit (self-reports) measures of personality to be 0.24, with about half of variability being due to moderating variables.

The promise of the IAT is that it should be less susceptible to faking. However, preliminary findings demonstrate the IAT is to a certain extent fakeable (Fiedler, Messner, & Bluemke, 2006). Given that there is still controversy about what the IAT measures (Rothermund & Wentura, 2004), and that there is a lot of method-specific (construct-irrelevant) variance associated with IATs (Mierke & Klauer, 2003), it is clear that more research is needed before IATs (and the related Go-No Go Association Test, Nosek & Banaji, 2001) can be regarded as viable tools in various applied educational contexts.

Implicit Assessments: Conditional Reasoning Tests (CRTs)

Conditional reasoning tests are multiple-choice tests consisting of items that look like reading comprehension or logical reasoning items, but are used to measure world-view, personality, biases, and motives (James, 1998; LeBreton, Barksdale, & Robin, 2007). Following a passage and a question, the CRT presents two or three logically incorrect alternatives, and two logically correct alternatives, which reflect different (often opposing) worldviews. Participants are asked to state which of the alternatives seems to be most reasonable, based on the information given in the text. Thus, respondents assume that they can solve a problem by reasoning about it, not realizing that there are two correct answers, and that their selection is guided by implicit assumptions underlying answer alternatives.

Participants are prompted to select one of the logically correct alternatives, presumably according to his or her underlying beliefs, rationalizing the selection through the use of justification mechanisms. For example, the examinee might select an

aggressive response to a situation, justifying it as an act of self-defense or as retaliation (LeBreton et al., 2007). These justification mechanisms serve to reveal hidden or implicit elements of the personality. To see an illustration of this idea, consider the example from LeBreton et al. (2007) presented in Tables 33.1–33.4.

The CRT for aggression has been shown to be unrelated to cognitive ability, yet reliable and valid for predicting different behavioral manifestations of aggression in the workplace (average r over 10 studies = 0.44) (James, McIntyre, & Glisson, 2004). Most of the research on CRTs has been in measuring aggression or achievement motivation (James, 1998). However, the method has proven difficult to replicate (Gustafson, 2004), and so far there is a paucity of research with children and adolescents (indeed, the cognitive difficulty of the items is currently such that applications would need to be restricted to high school populations and above). Also, as with IATs, the promise of resistance to faking has not been established (LeBreton et al., 2007). Thus, it seems that CRTs may need further work before being used in high-stakes academic situations. A sample CRT item is presented in Table 33.8.

Forced Choice

Peabody's (1967) early musings on personality assessments proposed a distinction between descriptive and evaluative judgements of personality. In any personality item, there is both a descriptive element and an evaluative element. For example, the descriptive element of the item "I am lazy" is that it measures conscientiousness (reverse-coded). The evaluative element is that "I am lazy" sounds like a bad thing. The basis for forced-choice testing is that test-takers are forced to choose between two or more statements that are equal in evaluative content (i.e., are equally socially desirable) but differ in terms of their descriptive content (i.e., measure different personality traits). Test-takers cannot then

Table 33.8 Conditional Reasoning Item (after James et al., 2004)

Item	Response
Half of all marriages end in divorce. One reason for the large number of divorces is that getting a divorce is quick and easy. If a couple can agree on how to split their property fairly, then they can get a divorce simply by filling out forms and taking them to court. They do not need lawyers. Which of the following is the most reasonable conclusion, based on the above?	(A) People are getting older when they get married. (B) If one's spouse hires a lawyer, then he or she is not planning to play fair. (C) Couples might get back together if getting a divorce took longer. (D) More men than women get divorced.

grade themselves highly on all positive statements, but must choose between them. Thus, faking-related variation in scores should be minimized.

There are several methods for forced-choice measurement, including pair comparisons, rank-ordering, and multidimensional forced-choice. In pair comparisons, the test-taker must choose between two equally desirable statements (e.g., "Which is more like you: 'I work hard' or 'I think up new ideas?'"). In rank-ordering, test-takers must rank a series of equally desirable statements in order from "most like me" to "least like me." Both these methods require that statements included in any one item be carefully matched for social desirability so that test-takers cannot use the evaluative aspects of the statements in their responses. In multidimensional forced-choice assessments, test-takers are presented with a dichotomous quartet of four different traits in which two socially desirable statements are paired with two socially undesirable statements (Jackson, Wroblewski, & Ashton, 2000). For example, a test-taker would be asked to select which is "most like you" and which is "least like you" from the following four statements: (1) "I work hard," (2) "I lose my temper," (3) "I love to help others," and (4) "I cannot deal with change." The statement selected as "most like you" would be scored +1; the statement selected as "least like you" would be scored -1, and the statements that were not endorsed at all would be scored zero.

There is some evidence to suggest that forced-choice tests are less fakeable than standard rating scales, and show stronger relationships with performance (e.g., Jackson et al., 2000; Martin, Bowen, & Hunt, 2002). However, forced-choice measures may have ipsative or partly ipsative properties. That is, scores on forced-choice measures may be appropriate for comparing the relative level of different traits within an individual, but inappropriate for comparing the relative levels of a trait across different people. Essentially, personality dimensions are not independent: one cannot be high on them all. This poses a problem for test-takers who really are high on multiple personality dimensions, or for test users who want to select individuals based on high scores on more than one personality dimension.

However, Stark, Chernyshenko, and Drasgow (2011) propose a number of IRT-based processes for constructing forced-choice items that ameliorate these issues of ipsativity. For example, in the sequential approach to developing a multi-dimensional pair-wise preference (MDPP) measure, one first determines both social desirability and item

parameters of a large number of items presented in conventional format. Social desirability ratings and item parameters may then be used to develop pairs of statements that act as a pair-comparison judgement (e.g., "I work hard" versus "I think up new ideas"). An important feature is that within most pair comparisons, items are drawn from two different personality domains (e.g., a Conscientiousness item is compared to an Agreeableness item). However, at least for some pairs, the items belong to the same domain. That way, it is possible to compute scores that also have a normative value; that is, can be used to differentiate between people. Empirical evidence to date suggests that tests constructed and scored using the MDPP method appear resistant to faking, and that normative rather than ipsative information can be recovered from this process. In this way, an empirically based procedure for item selection and test development combined with new statistical modeling techniques seems to produce the best of both worlds: fake-proof tests that also lack the ipsativity that plagued earlier operationalizations of forced-choice measurement. For this reason, we contend the approach will be explored much more in educational contexts in the years ahead. For example, an upcoming Program for International Student Assessment (PISA) field trial uses several variants of this approach to assess learning strategies. Table 33.9 contains a sample multidimensional forced-choice item.

Bayesian Truth Serum

The Bayesian Truth Serum (BTS) calculates how often people endorse item content they perceive as unusual and therefore possibly undesirable (Prelec, 2004). A person who rarely agrees with unpopular attitudes or behaviors is assumed to be adjusting their answers based on conformity to group opinion or expectations, rather than giving an honest appraisal of the item content. The extent of agreement with self-perceived "unpopular" items can be taken as an index of truth-telling. The method requires the test-taker to provide two pieces of

Table 33.9 Multi-dimensional Forced-Choice Item

Item	Response
Consider the four statements at right.	I work hard. I lose my temper.
Which is most like you, and which is least like you?	I love to help others. I cannot deal with change.

information for every item: their own response, and the proportion of people they estimate would respond the same way on that item. For example, a test-taker might have to choose which statement described them better: "I think up new ideas" or "I work hard." The test-taker would also have to estimate the preference of the wider population (e.g., guessing that perhaps 75% of the overall population would choose "I think up new ideas" over "I work hard" to describe themselves). Prelec argues that respondents who answer questions honestly will tend to overestimate the percentage of other people who agree with them. In this way, an index of truth telling can be calculated.

However, estimating the beliefs of other people is both meta-cognitively complex, and may be subject to frame-of-reference effects as to who these "others" are considered to be (other college applicants, other people the test-taker personally knows, other people of similar SES and demographics to the test-taker, or all other people in the world). For these reasons, the BTS may not function accurately for test-takers with poor meta-cognitive skills or test-takers using an unusual frame of reference. In addition, collecting additional information doubles the test-taking time, since twice as many questions are asked. Like the conditional reasoning test, the complexity of the Bayesian truth serum suggests that the method might only be accurately applied after the mid-teens, and then only among cognitively normal populations. Young children may not have the meta-cognitive skill to answer such questions, nor the concentration to sit through tests of double the normal length.

Future Directions: Potential Uses of Non-cognitive Assessments

In the final section of this chapter, we consider the findings from the literature analysis, and synthesize them in a way to suggest several ways in which comprehensive non-cognitive assessments might be used (or are currently used) in education.

High-Stakes Assessment

High-stakes applications of non-cognitive tests in education include diagnosis and selection. For diagnosis, non-cognitive assessments have an important role to play in augmenting traditional cognitive assessments aimed at diagnosing learning disorders and difficulties. For example, test anxiety shows frequent comorbidity with learning difficulties, as students who struggle with scholastic material develop an aversion to and fear of evaluative

situations (e.g., Bryan, Sonnefeld, & Grabowski, 1983; Sena, Lowe, & Lee, 2007). Test anxiety can function as both a symptom of learning difficulties and an exacerbation of learning difficulties, and may even affect scores on cognitive assessments designed to diagnose learning difficulties. Thus, this particular non-cognitive factor may be a pertinent factor to consider when diagnosing a student with a particular learning disorder.

The second major high-stakes application of non-cognitive assessments is selection for college, graduate school, preparatory school, gifted classes, or the honors or advanced placement track. Large-scale, high-stakes non-cognitive assessment, based on faculty ratings, has been implemented recently for graduate school admissions (Kyllonen, 2008). If this is successful, it is reasonable to expect that a similar application for undergraduate admissions could follow. Recent meta-analytic evidence shows that non-cognitive assessments are just as predictive of academic achievement as traditional intelligence testing (Poropat, 2009). Moreover, this prediction is separate from intelligence, that such non-cognitive assessments retain their strong relationship to achievement even after controlling for ability. Such results imply that accuracy in selection decisions could be improved drastically with the addition of non-cognitive assessments. For example, two students of equal intelligence might have very different chances of success in advanced placement classes if one of them is willing to work hard, seek help, and extend effort to maintain a supportive social network, whereas the other believes that success is due to lucky breaks and chance factors. Using appropriate non-cognitive tests to augment selection decision could feasibly result in better outcomes. Moreover, there is some evidence to suggest that including non-cognitive factors in selection decisions would result in less adverse impact on ethnic minorities and women (e.g., McDaniel et al., 2001).

A diagnosis or selection decision may be vastly influential for children and adolescents, with long-term consequences. Therefore, the potential for motivated individuals to try to gain a particular outcome should be not ignored: it is quite possible that even young children will give a socially correct answer rather than one that describes their actual tendencies. For this reason, several safeguards against response distortion should be considered when tests are used for high-stakes purposes. First, a non-cognitive assessment should never be the sole basis for a high-stakes decision, but should be

considered in conjunction with other key factors (e.g., cognitive test scores, teacher and parent interviews, and record of achievement relative to ability for learning disorder diagnosis). Second, in most selection scenarios, it may be more beneficial to exclude those with unacceptably low non-cognitive skills, rather than select the top few with exceptionally high non-cognitive skills. Third, an aggregate of multiple other-reports should be preferred over self-reports, as in the case of the Personal Potentiality Index assessment for post-graduate admissions (Kyllonen, 2008). With these appropriate safeguards in mind, non-cognitive assessments have the potential to drastically improve selection and diagnosis decisions.

Developmental Scales

Another possible application of non-cognitive assessments is to track students' development over time (Roberts, 2009; Roberts & Wood, 2006). At the individual level, a "report card" each year could show the students' development of non-cognitive skills such as impulse control, social skills, coping strategies, and attitudes towards school and school-work. Particularly in the early grades, this sort of feedback could feasibly be used for early identification of individuals or cohorts at risk for learning disorders, conduct disorders, or other academic or social problems. At the institutional level, schools that provide programs for social and emotional learning, peer support, or the development of academic skills such as time management or learning strategies could monitor the progress of their student body in developing and maintaining these skills. Schools would also be able to track trends at specific levels (e.g., the adjustment to a new environment by students beginning kindergarten or high school; or the stress experienced by students undertaking college preparations in the last two years of secondary school). At the wider level, district, state or national comparisons of students' non-cognitive tests scores would allow a strong evidence base for policy development in education.

Interventions

One of the strengths of using non-cognitive tests as developmental scales is the potential for interventions or training. There are several ways that assessments can form the backbone of intervention and training development. First, as mentioned above, large-scale developmental assessments can be used to guide policy change, as well as to evaluate the effectiveness of policy implementation. Second, scores

on specific assessments may serve as the basis for particular suggestions in the form of tailored feedback and action plans. For example, a high-school student might complete a three-component measure of time management, and score much lower on the *planning* component than the other two parts. Such a student would receive the information that they are (for example) good at coping with change, and tend to refrain from procrastination, but that planning is their weak point. Such information should increase their self-knowledge (and potentially their meta-cognitive skills). The student could also receive a series of suggestions about how to improve their planning skills, some helpful tools such as a weekly or monthly time-and-task calculator, or a referral to existing school- or district-based programs for assisting with academic-readiness skills. As we mentioned in earlier sections of this chapter, the TpB is a particularly powerful technique for linking assessments with interventions, as its theoretical basis focuses on behavior and behavior change (Ajzen, 2011).

Educational applications of non-cognitive testing might also borrow from personnel psychology, applying ideas such as the development assessment center, which directly links assessment with development activities (e.g., Thornton & Rupp, 2005). In developmental assessment centers, the test-taker:

- (a) learns about the non-cognitive dimensions the test measures (e.g., learns about the underlying components of time management);
- (b) learns their own strengths and weaknesses;
- (c) learns how to set goals to improve;
- (d) learns how to monitor their progress in improvement;
- (e) is provided with exercises, feedback, and experiential learning activities.

These programs are currently being implemented by several research teams, and the initial results are promising (Elias & Clubby, 1992). This paradigm holds great promise for improving non-cognitive skills for education.

Concluding Comments

In this chapter we presented an overview of a rather wide variety of both conventional and novel methods for assessing non-cognitive skills in an educational context. Self-assessments are the most common and are likely to be useful in any kind of non-cognitive assessment system, particularly when the stakes are not high. Other-ratings, such as teacher ratings, parent reports, letters of recommendation, and interviews, are also quite useful, and as

discussed in corresponding sections of this chapter, they are currently the most viable for high-stakes selection applications. A range of non-traditional assessments hereby reviewed—such as the implicit association test, day reconstruction method, and conditional reasoning tests—are intriguing and may potentially be quite useful for assessing students' non-cognitive characteristics. Situational judgement tests are an increasingly popular way to measure non-cognitive characteristics. They have been used in so many studies over the past 10 years that the methodology for developing them is now fairly affordable, and the measures are becoming increasingly reliable and valid. All of these methods are constantly evolving, so more information attesting to their validity and applicability for educational contexts will be accrued in the upcoming years.

Overall, our understanding of non-cognitive factors influencing academic achievement and possible approaches toward the measurement and assessment of these factors allows us to identify students who are more or less likely to do well in a specific academic program. Additionally, our knowledge of the relationships among a range of non-cognitive constructs and educational outcomes can be used to develop effective interventions. These interventions can be successful in enhancing students' non-cognitive characteristics, and, consequently, their achievement. In sum, these constructs can be successfully assessed and modified, and it is our hope that more and more researchers will start designing studies investigating the quality of such assessments and the effectiveness of such interventions.

Authors' Note

We thank Anthony Betancourt, Jeremy Burrus, Daniel Howard, Teresa Jackson, Stefan Krumm, Mary Lucas, Bobby Naemi, Jen Minsky, and Patrick Kyllonen for supporting the preparation of this manuscript. All statements expressed in this chapter are the authors' and do not reflect the official opinions or policies of the authors' host affiliations. Correspondence concerning this article should be directed to Anastasiya A. Lipnevich via email: a.lipnevich@gmail.com.

References

- Abe, J. A. A. (2005). The predictive value of the Five-Factor Model of personality with preschool-age children: A nine year follow-up study. *Journal of Research in Personality*, 39, 423–442.
- Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist 4–18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior & Human Decision Processes*, 50, 179–211.
- Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, 32, 665–683.
- Ajzen, I. (2011). Behavioral interventions: Design and evaluation guided by the theory of planned behavior. In M. M. Mark, S. I. Donaldson, & B. C. Campbell (Eds.), *Social psychology for program and policy evaluation* (pp. 74–100). New York: Guilford.
- Andrews, J. A., Hampson, S. E., Greenwald, A. G., Gordon, J., & Widdop, C. (2010). Using the Implicit Association Test to assess children's implicit attitudes toward smoking. *Journal of Applied Social Psychology*, 40, 2387–2406.
- Armitage, C., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology*, 40, 471–499.
- Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, 86, 736–765.
- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, 35, 281–322.
- Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality & Individual Differences*, 40, 1235–1245.
- Baird, L. L., & Knapp, J. E. (1981). *The inventory of documented accomplishments for graduate admissions: Results of a field trial study and its reliability, short-term correlates, and evaluation*. (ETS Research Rep. No. 81–18, GRE Board Research Rep. No. 78–3R). Princeton, NJ: Educational Testing Service.
- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational & Psychological Measurement*, 60, 361–370.
- Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, 81, 261–272.
- Belli, R. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, 6, 383–406.
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection & Assessment*, 14, 317–335.
- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78, 246–263.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, 117, 187–215.
- Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality & Social Psychology*, 79, 631–643.
- Bryan, J. H., Sonnetfeld, J. L., & Grabowski, B. (1983). The relationship between fear of failure and learning disabilities. *Learning Disability Quarterly*, 6, 217–222.

- Burrus, J., MacCann, C., Kyllonen, P., & Roberts, R. D. (2011). Non-cognitive constructs in K-16: Assessments, interventions, educational and policy implications. In P. J. Bowman & E. P. St John (Eds.), *Diversity, merit, and higher education: Toward a comprehensive agenda for the twenty-first century* (pp. 233-274). Ann Arbor, MI: AMS Press.
- Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection & Assessment*, 8, 248-260.
- Campbell, J. R., Voelkl, K. E., & Donahue, P. L. (1997). *NAEP 1996 trends in academic progress*. (NCES Publication No. 97985r). Washington, DC: U.S. Department of Education.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgement tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143-159.
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgement tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.
- Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343-359). Mahwah, NJ: Erlbaum.
- Connell, J. P., Spencer, M. B., & Aber, J. L. (1994). Educational risk and resilience in African-American youth: Context, self, action, and outcomes in school. *Child Development*, 65, 493-506.
- Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection & Assessment*, 16, 155-169.
- Costa, P. T. Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, 64, 21-50.
- Crede, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, 3, 425-453.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349-354.
- Davis, L. E., Ajzen, I., Saunders, J., & Williams, T. (2002). The decision of African American students to complete high school: An application of the theory of planned behavior. *Journal of Educational Psychology*, 94, 810-819.
- Delgallarrando, M. G. (July 9, 2008). *Validan plan de admisión complementaria a la UC*. (p. 9), El Mercurio, Santiago Chile.
- Digman, J. M., & Inouye, J. (1986). Further specification of the five robust factors of personality. *Journal of Personality & Social Psychology*, 50, 116-123.
- Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments? *Psychology Science*, 48, 209-225.
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, 13, 440-484.
- Duckworth, A., & Seligman, M. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16, 939-944.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256-273.
- Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, 16, 1-23.
- Elias, M. J., & Clabby, J. (1992). *Building social problem solving skills: Guidelines from a school-based program*. San Francisco: Jossey-Bass.
- Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology*, 84, 155-166.
- Etienne, P. M., & Julian, E. R. (2005). Assessing the personal characteristics of premedical students. In W. J. Camara & E. W. Kimmel (Eds.), *Choosing students: Higher education admissions tools for the 21st century* (pp. 215-230). Mahwah, NJ: Erlbaum.
- Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I," the "A," and the "T": A logical and psychometric critique of the Implicit Association Test. (IAT). *European Review of Social Psychology*, 17, 74-147.
- Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal & Social Psychology*, 44, 329-344.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34, 906-911.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132, 692-731.
- Goho, J., & Blackman, A. (2006). The effectiveness of academic admission interviews: an exploratory meta-analysis. *Medical Teacher*, 28, 335-340.
- Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality & Social Psychology*, 85, 541-553.
- 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, 1464-1480.
- Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality & Social Psychology*, 79, 1022-1038.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3-25.
- Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist*, 61, 56-61.
- Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, 36, 341-357.
- Gustafson, S. (Chair). (2004). *Making conditional reasoning test work: Reports from the frontier*. Symposium conducted at the 19th Annual Conference of the Society for Industrial and Organizational Psychology Conference, Chicago, IL.
- Hancock, J. T., Curry, L., Gootah, S., & Woodworth, M. T. (2004). Lies in conversation: An examination of deception using automated linguistic analysis. *Proceedings, Annual Conference of the Cognitive Science Society*, 26, 534-540.

- Hardeman, W., Johnston, M., Johnston, D. W., Bonetti, B., Wareham, N. J., & Kinmonth, A. L. (2002). Application of the Theory of Planned Behaviour in behaviour change interventions: a systematic review. *Psychology & Health, 17*, 123-158.
- Harzing, A.-W. (2006). Response styles in cross-national survey research. *International Journal of Cross-Cultural Management, 6*, 243-266.
- Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.
- Hell, B., Trapmann, S., Weigand, S., & Schuler, H. (2007). Die Validität von Auswahlgesprächen im Rahmen der Hochschulzulassung—eine Metaanalyse. *Psychologische Rundschau, 58*, 93-102.
- Hendriks, A. A. J., Kuyper, H., Offringa, J. G., & Van der Werf, M. P. (2008). Assessing young adolescents' personality with the five-factor personality inventory. *Assessment, 15*, 304-316.
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality & Social Psychology Bulletin, 31*, 1369-1385.
- Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection & Assessment, 12*, 262-273.
- Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*, 371-388.
- 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*, 501-511.
- James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*, 131-163.
- James, L. R., McIntyre, M. D., & Glisson, C. A. (2004). The Conditional Reasoning Measurement System for aggression: An overview. *Human Performance, 17*, 271-295.
- Janz, T., Hellervik, L., & Gillmore, D. C. (1986). *Behavior description interview*. Boston, MA: Allyn & Bacon.
- John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*, 521-551.
- Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science, 306*, 1776-1780.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford Press.
- Kim, S., & Kyllonen, P. C. (2008). Rasch measurement in developing faculty ratings of students applying to graduate school. *Journal of Applied Measurement, 9*, 168-181.
- Krokos, K. J., Meade, A. W., Cantwell, A. R., Pond, S. B., & Wilson, M. A. (2004). Empirical keying of situational judgment tests: Rationale and some examples. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
- Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). Attitude measurement. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change*. Mahwah, NJ: Erlbaum.
- Kyllonen, P. C. (2008). *The research behind the ETS Personal Potential Index*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Products/PPI/10411_PPI_bkgnd_report_RD4.pdf on January 15, 2009.
- Kyllonen, P. C., & Kim, S. (2004). Personal qualities in higher education: Dimensionality of faculty ratings of graduate school applicants. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.
- Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence*. Thousand Oaks, CA: Sage Publications, Inc.
- Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.
- LeBreton, J. M., Barksdale, C. D., & Robin, J. (2007). Measurement issues associated with Conditional Reasoning Tests: Indirect measurement and test faking. *Journal of Applied Psychology, 92*, 1-16.
- Li, A., & Bagger, J. (2006). Using the BID-R to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection & Assessment, 14*, 131-141.
- Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection & Assessment, 10*, 245-257.
- Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgement tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*, 1181-1188.
- Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgement test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442-452.
- Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and non-cognitive ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*, 101-128.
- Lipnevich, A. A., MacCann, C., Krumm, S., & Roberts, R. D. (2011). Mathematics attitudes in Belarusian and U.S. middle school students. *Journal of Educational Psychology, 103*, 105-118.
- Liu, O. L. (2009). Evaluation of a learning strategies scale for middle school students. *Journal of Psychoeducational Assessment, 27*, 213-322.
- Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2007). *The standardized letter of recommendation: Implications for selection*. ETS Research Report RR-07-38. Princeton, NJ: ETS.
- Liu, O. L., Rijmen, F., MacCann, C., & Roberts, R. D. (2009). The assessment of time management in middle-school students. *Personality & Individual Differences, 47*, 174-179.
- MacCann, C., Duckworth, A., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning & Individual Differences, 19*, 451-458.
- MacCann, C., Fogarty, G. J., & Roberts, R. D. (2012). Strategies for success in vocational education: Time management

- is more important for part-time than full-time students. *Learning & Individual Differences*, 22, 518–623.
- MacCann, C., Lipnevich, A. A., & Roberts, R. D. (2013). Parents know best! Comparing self- and parent-reported personality in predicting academic achievement. Submitted to the ETS Research Report Series. Princeton, NJ: ETS.
- MacCann, C., Matthews, G., Wang, L., & Roberts, R. D. (2010). Emotional intelligence and the eye of the beholder: Comparing self- and parent-rated situational judgements in adolescents. *Journal of Research in Personality*, 44, 673–676.
- MacCann, C., Ziegler, M., & Roberts, R. D. (2011). Faking in personality assessment: Reflections and recommendations. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 309–329). New York: Oxford University Press.
- Marchese, M. C., & Muchinski, P. M. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection & Assessment*, 1, 18–26.
- Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology*, 23, 498–509.
- Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, 80, 366–380.
- Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality & Individual Differences*, 32, 247–256.
- McDaniel, M. A., Morgesen, 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, 730–740.
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599–616.
- McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. J. Rumsey, C. D. Walker, & J. Harris (Eds.), *Personnel selection and classification research* (pp. 193–232). Mahwah, NJ: Lawrence Erlbaum.
- Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality & Social Psychology*, 84, 857–870.
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, 56, 128–165.
- Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality & Social Psychology*, 85, 1180–1192.
- Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology*, 79, 272–280.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality & Social Psychology Bulletin*, 29, 665–675.
- Norman, W. T. (1963). Personality measurement, faking, and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology*, 47, 225–241.
- Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, 19, 161–176.
- Olson-Buchanan, J. B., & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 253–278). San Francisco: Jossey-Bass.
- Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology*, 51, 1–24.
- Oltmanns, T. F., & Turkheimer, E. (2006). Perceptions of self and others regarding pathological personality traits. In R. Krueger & J. Tackett (Eds.), *Personality and psychopathology: Building bridges*. New York: Guilford.
- Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). The role of social desirability in personality testing in personnel selection: The red herring. *Journal of Applied Psychology*, 81, 660–679.
- Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgement inventory as predictors of college student performance. *Journal of Applied Psychology*, 89, 187–207.
- Paulhus, D. L. (1998). *The Balanced Inventory of Desirable Responding (BIDR-7)*. Toronto/Buffalo: Multi-Health Systems.
- Pauonen, S. V., & Ashton, M. C. (2001). Big five predictors of academic achievement. *Journal of Research in Personality*, 35, 78–90.
- Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality & Social Psychology*, 7, 1–13.
- Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology*, 98, 583–597.
- Poropat, A. E. (2009). A meta-analysis of the Five-Factor Model of personality and academic performance. *Psychological Bulletin*, 135, 322–338.
- Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306, 462–466.
- Prelec, D., & Weaver, R. G. (2006). Truthful answers are surprisingly common: Experimental tests of Bayesian truth serum. Paper presented at the ETS Mini-conference on Faking in Non-cognitive Assessments. Princeton, NJ: Educational Testing Service.
- Rammstedt, B., & Krebs, D. (2007). Does response scale format affect the answering of personality scales? Assessing the big five dimensions of personality with different response scales in a dependent sample. *European Journal of Psychological Assessment*, 23, 32–38.
- Reynolds, C. R., & Kamphaus, R. W. (2004). *BASC-2: Behavior Assessment System for Children, second edition manual*. Circle Pines, MN: American Guidance Service.
- Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, 85, 880–887.
- Robbins, S., Lauver, K. J., 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.
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality*, 43, 137–145.
- Roberts, B. W., & Wood, D. (2006). Personality development in the context of the Neo-socioanalytic model of personality.

- In D. Mroczek & T. Little (Eds.), *Handbook of personality development* (pp. 11–39). Mahwah, NJ: Lawrence Erlbaum Associates.
- Roberts, R. D., Berancourt, A. C., Burrus, J., Holtzman, S., Libbrecht, N., MacCann, C., et al. (2011). *Multimedia assessment of emotional abilities: Development and validation*. Army Research Institute Report Series. Arlington, VA: ARI.
- Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, 133, 139–165.
- Rutter, D. (2000). Attendance and reattendance for breast cancer screening: a prospective 3-year test of the Theory of Planned Behaviour. *British Journal of Health Psychology*, 5, 1–13.
- Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology*, 52, 359–391.
- Sackett, P. R. (2006). *Faking and coaching effects on non-cognitive predictors*. Paper presented at the ETS Mini-conference on Faking in Non-cognitive Assessments. Princeton, NJ: Educational Testing Service.
- Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality & Social Psychology*, 46, 929–938.
- 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, 262–274.
- Schuler, H. (2002). *Das Einstellungsinterview*. Göttingen, Germany: Hogrefe.
- Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research*, 4, 27–41.
- Sena, J. D. W., Lowe, P. A., & Lee, S. W. (2007). Significant predictors of test anxiety among students with and without learning disabilities. *Journal of Learning Disabilities*, 40, 360–376.
- Sheeran, P. (2002). Intention-behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, 12, 1–36.
- Small, E. E., & Diefendorff, J. M. (2006). The impact of contextual self-ratings and observer ratings of personality on the personality-performance relationship. *Journal of Applied Social Psychology*, 36, 297–320.
- Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology*, 100, 961–976.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement*, 29, 184–201.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2011). Constructing fake-resistant personality tests using Item Response Theory: High stakes personality testing with Multidimensional Pairwise Preferences. In M. Zeigler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 214–239). New York: Oxford University Press.
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., & Williams, W. M. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.
- Stricker, L. J., Rock, D. A., & Bennett, R. E. (2001). Sex and ethnic-group differences on accomplishment measures. *Applied Measurement in Education*, 14, 205–218.
- Thornton, G. C. III, & Rupp, D. E. (2005). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.
- Trapmann, S., Hell, B., Hirn, J. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Journal of Psychology*, 215, 132–151.
- Tupes, E. C., & Christal, R. E. (1961/1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, 60, 225–251.
- Vannelli, J., Kuncel, N. R., & Ones, D. S. (2007, April). A mixed recommendation for letters of recommendation. In N. R. Kuncel (Chair), *Alternative Predictors of Academic Performance: The Glass is Half Empty*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fake-ability estimates: Implications for personality measurement. *Educational & Psychological Measurement*, 59, 197–210.
- Wagerman, S. A., & Funder, D. C. (2006). Acquaintance reports of personality and academic achievement: A case for conscientiousness. *Journal of Research in Personality*, 41, 221–229.
- Walters, A., Kyllonen, P. C., & Plante, J. W. (2006). Developing a standardized letter of recommendation. *The Journal of College Admission*, 191, 8–17.
- Walters, A. M., Kyllonen, P. C., & Plante, J. W. (2003). Preliminary research to develop a standardized letter of recommendation. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multi-method approach. *Canadian Journal of School Psychology*, 24, 108–124.
- Waugh, G. W., & Russell, T. L. (2003). Scoring both judgement and personality in a situational judgement test. Paper presented at the 45th Annual Conference of the International Military Testing Association. Pensacola, Florida.
- Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*, 50, 25–49.
- Wilkinson, D., Zhang, J., Byrne, G. J., Luke, H., Ozolins, I. Z., Parker, M. H., et al. (2008). Medical school selection criteria and the prediction of academic performance: Evidence leading to change in policy and practice at the University of Queensland. *Medical Education*, 188, 349–354.
- Wine, J. (1971). Test anxiety and the direction of attention. *Psychological Bulletin*, 76, 92–104.
- Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum Press.
- Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, 7, 168–190.
- Ziegler, M., Schmidt-Atzert, L., Bühner, M., & Krumm, S. (2008). Motivated, unmotivated, or faking? Susceptibility of three achievement motivation tests measures to faking: Questionnaire, semi-projective, and objective. *Psychology Science*, 49, 291–307.