Raps, C. S. (2009). When we research "conditions–treatments," is there a misnomer? *American Psychologist, 64,* 275–276.
Silverstein, L. B., & Auerbach, C. F. (2009). Using qualitative research to develop culturally competent evidence-based practice. *American Psychologist, 64,* 274–275.
Tryon, W. W. (2009). Missing mechanism information. *American Psychologist, 64,* 273–274.
Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition.* Stanford, CA: Stanford University Press.

Correspondence concerning this comment should be addressed to Alan E. Kazdin, Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520-8205. E-mail: alan.kazdin@yale.edu

DOI: 10.1037/a0014923

(Over and) Beyond High-Stakes Testing

Angela L. Duckworth
University of Pennsylvania

Sackett, Borneman, and Connelly's (May–June 2008) article and recent meta-analyses (e.g., Kuncel & Hezlett, 2007) should lay to rest any doubt over whether high-stakes standardized tests predict important academic and professional outcomes—they do. The challenge now is to identify *noncognitive* individual differences that determine the same outcomes.

Noncognitive is, of course, a misnomer. Every psychological process is cognitive in the sense of relying on the processing of information of some kind. Moreover, characteristic patterns of attending to and interpreting information underlie many if not most personality traits (Bandura, 1999; Mischel & Shoda, 1999). Finally, emotion and personality certainly influence the quality of one's thinking (Baron, 1982).

Why do so many psychologists, including myself, resort to the term *noncognitive* despite its obvious inappropriateness? My guess is that most of us use the term *cognitive* in this context as shorthand for cognitive ability and knowledge, constructs validly measured by high-stakes tests. Convinced that such tests explain a significant proportion of variance in performance, we are interested in understanding what *else* predicts achievement. We want to know, for instance, why some graduate students or employees are dramatically more successful than others, even after controlling for SAT scores. Used this way, the term *noncognitive* embraces personality and temperament traits, interests, values, and goals.

At first glance, it may seem that what distinguishes cognitive from noncognitive traits is that the former describe what one *can* do and the latter what one *usually* does. It is true that cognitive ability and knowledge are capacities that depend on the motivation of the individual to be expressed; they are upper bounds on behavior. It is also true that many noncognitive traits (e.g., sociability, kindness) are, in contrast, propensities, or patterns of behavior; they are the expected value (i.e., mean) of behavior.

It is a mistake to conflate the cognitive versus noncognitive dimension with the capacity versus propensity dimension. See Table 1. For any given capacity, there is an accompanying propensity. The propensity to act in a certain way is obviously limited by one's capacity to act in that way, but propensity and capacity are surprisingly weakly correlated (Sackett, Zedeck, & Fogli, 1988). For instance, the tendency to put forth intellectual effort in day-to-day living is not impressively correlated with measures of fluid intelligence (Ackerman & Heggestad, 1997).

Whereas intelligence is almost always studied as a capacity rather than a propensity, the opposite is true of noncognitive traits. In fact, for most noncognitive constructs, valid measures of capacity have not been developed (e.g., how sociable an individual could act when maximally motivated to do so). One exception is self-control. As a capacity, self-control has been operationally defined as the number of seconds preschoolers can wait for a preferred, larger treat over a smaller one. However, particularly among older subjects, self-control is more often assessed using questionnaire items such as "I refuse things that are bad for me." In this case, what is being measured is a propensity—specifically, the level of self-control an individual demonstrates in daily life.

An important feature of capacities, whether cognitive or noncognitive, is that they are amenable to measurement by performance tasks in which motivation is maximized (e.g., high-stakes tests). Propensities, in contrast, are only measurable by integrating observations of an individual's behavior over long periods of time. Who is in a position to make such observations? Only informants and the individual himself, all of whose judgments are subjective.

I suspect that the necessarily subjective nature of questionnaires has made this mode of assessment unattractive to policymakers and testing companies. In contrast to intelligence and achievement tests, the vast majority of questionnaires used in psychology research lack current norms based on large samples. Lack of information on normative variance in the population makes it impossible to follow Sackett et al.'s (2008) advice to use psychometric formulas to correct estimates of predictive validity for restriction on range.

Consider the study by Duckworth and Seligman (2005), which Sackett and colleagues (2008) used to demonstrate the importance of correcting for range restriction. In this study, both self-control and IQ were used to predict grade point average and other outcomes among adolescents at a public magnet school. Admission criteria

Table 1
Two Dimensions Distinguishing Individual Differences That Predict Achievement

	Capacity–propensity dimension	
Cognitive–noncognitive dimension	Capacity	Propensity
Cognitive (i.e., related to thinking and reasoning)	Intelligence (IQ tests), knowledge (achievement tests)	Typical intellectual engagement
Noncognitive (i.e., not related to thinking and reasoning)	Self-control measured as the ability to delay gratification	Self-control measured by self-report or informant ratings of behavior

to this school included report card grades and achievement test scores—but not IQ scores. In the study sample, both grades and achievement test scores were highly correlated with self-control and, to a lesser extent, IQ. Thus, there is every reason to suspect that restriction on range was at least as severe for self-control as it was for IQ. The IQ measure used, the Otis Lennon School Ability Test, has published norms based on data collected in 1995 from over 450,000 students in over 10,000 American school districts. It was therefore possible to correct estimates of the predictive validity of IQ for restriction on range. In contrast, normative data were available for only one of several measures of self-control: A sample of 230 British 13-year-olds completed the Eysenck Junior Impulsiveness subscale in 1984. (I note also that Sackett et al.'s excellent review misreported the direction of the association between intelligence and procrastination in this study. Self-controlled students procrastinated less, whereas more intelligent students procrastinated more.)

In which direction does progress lie? For the particular problem of range restriction, one might exploit more sophisticated means of correcting parameter estimates. For instance, in the econometric literature, the two-stage Heckman correction is widely used to correct for bias in estimates derived from selected samples. The Heckman correction has never been used in a published psychology article.

Coincidentally, since winning the Nobel Prize for his work on selection bias, Heckman has devoted himself to the study of noncognitive individual differences and their effects on wages and other important outcomes (see Borghans, Duckworth, Heckman, & ter Weel, 2008). Given that a shared goal of the social sciences is to use the most powerful measurement and analysis tools to understand and improve the human condition, it seems that in exploring the vast territory of individual differences not captured by high-stakes tests, cross-disciplinary collaboration may be the most promising way forward.

REFERENCES

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219–245.

Bandura, A. (1999). Social cognitive theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 154–196). New York: Guilford Press.

Baron, J. (1982). Personality and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 308–351). New York: Cambridge University Press.

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

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

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315,* 1080–1081.

Mischel, W., & Shoda, Y. (1999). Integrating dispositions and processing dynamics within a unified theory of personality: The cognitive-affective personality system. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 197–218). New York: Guilford Press.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63,* 215–227.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482–486.

Correspondence concerning this comment should be addressed to Angela L. Duckworth, Department of Psychology, University of Pennsylvania, 3701 Market Street, 209, Philadelphia, PA 19104. E-mail: duckworth@psych.upenn.edu

DOI: 10.1037/a0014926

Being Creative With the Predictors *and* Criteria for Success

James C. Kaufman and Mark D. Agars
California State University, San Bernardino

Sackett, Borneman, and Connelly (May–June 2008) argued that several common criticisms of cognitively laden tests are not well supported by the literature. The authors' systematic exploration of research surrounding seven specific criticisms is laudable, and we do not find fault with their conclusions as presented. In evaluating the seven concerns, however, the authors largely neglected the criteria that such tests are intended to predict. As a result, readers may come away with the erroneous conclusion that all is well in the mass testing world of cognitive ability. We wish to expand on Sackett et al.'s review by raising concerns about traditional approaches to defining academic and organizational success. In doing so, we argue for the importance of creativity.

The authors addressed the "popular" concerns, with a focus on the test-taking process and the measures themselves. Consequently, the larger decision-making context in which such tests are used was given little consideration. The strengths and weaknesses of any predictor are necessarily linked to the established criteria and cannot be adequately evaluated without also examining the appropriateness of the measured outcomes it is intended to predict. The limitations of cognitive testing, and the importance of alternative predictors, are best evaluated in the context of expanded criteria.

The debate about "alternative" predictors of performance has a long history in industrial/organizational psychology and education. Although cognitive ability remains firmly entrenched as the strongest predictor of performance (Schmidt & Hunter, 2004), other constructs (e.g., conscientiousness) have predictive power. However, less time has been spent examining criteria (Day & Schleicher, 2007). We argue that criterion deficiency and stagnation may be as plausible a reason for the continued validation of cognitive tests as the integrity of the measures themselves.

The increased importance of cognitive-based standardized tests in education and employment selection naturally limits the use of outcomes other than traditional measures (e.g., grade point average). Are grades the best possible criteria, however? Enright and Gitomer (1989) identified seven important generalized competencies that graduate school professors and admissions committees identified as being most important; the ability to get good grades was not on this list. Demonstrating the limitations of cognitive-based tests, Sternberg and Williams (1997) found some evidence that when the criteria were expanded to include dissertation ratings, the predictive power of Graduate Record Examination scores dropped. The danger is that these findings may be used as evidence that such outcomes are not relevant criteria rather than as an indication that cognitive-based predictors are not as comprehensive as we tend to believe. We believe we must also critically examine expanded criteria.

Indeed, Sackett et al. (2008) acknowledged these concerns when they noted, "It is important to differentiate between technical questions . . . and values-based questions, such as those about the *relative importance of one criterion versus another*" (p. 225, italics added). Clearly their article