Psychological Bulletin

How Smart Do You Think You Are? A Meta-Analysis on the Validity of Self-Estimates of Cognitive Ability

Philipp Alexander Freund and Nadine Kasten

Online First Publication, December 19, 2011. doi: 10.1037/a0026556

CITATION

Freund, P. A., & Kasten, N. (2011, December 19). How Smart Do You Think You Are? A Meta-Analysis on the Validity of Self-Estimates of Cognitive Ability. *Psychological Bulletin*. Advance online publication. doi: 10.1037/a0026556

Psychological Bulletin

0033-2909/11/$12.00 DOI: 10.1037/a0026556

How Smart Do You Think You Are? A Meta-Analysis on the Validity of Self-Estimates of Cognitive Ability

Philipp Alexander Freund
Leuphana University at Lüneburg

Nadine Kasten
University of Osnabrück

Individuals' perceptions of their own level of cognitive ability are expressed through self-estimates. They play an important role in a person's self-concept because they facilitate an understanding of how one's own abilities relate to those of others. People evaluate their own and other persons' abilities all the time, but self-estimates are also used in formal settings, such as, for instance, career counseling. We examine the relationship between self-estimated and psychometrically measured cognitive ability by conducting a random-effects, multilevel meta-analysis including a total of 154 effect sizes reported in 41 published studies. Moderator variables are specified in a mixed-effects model both at the level of the individual effect size and at the study level. The overall relationship is estimated at $r = .33$. There is significant heterogeneity at both levels (i.e., the true effect sizes vary within and between studies), and the results of the moderator analysis show that the validity of self-estimates is especially enhanced when relative scales with clearly specified comparison groups are used and when numerical ability is assessed rather than general cognitive ability. The assessment of less frequently considered dimensions of cognitive ability (e.g., reasoning speed) significantly decreases the magnitude of the relationship. From a theoretical perspective, Festinger's (1954) theory of social comparison and Lecky's (1945) theory of self-consistency receive empirical support. For practitioners, the assessment of self-estimates appears to provide diagnostic information about a person's self-concept that goes beyond a simple "test-and-tell" approach. This information is potentially relevant for career counselors, personnel recruiters, and teachers.

Keywords: self-estimates, self-concept, cognitive ability, meta-analysis

People assess their own abilities in various situations of everyday life. Although we may not be consciously aware of it, we frequently consider our physical and mental constitution when we encounter both novel and routine tasks (cf. Ackerman & Wolman, 2007). For instance, we might contemplate whether our abilities and skills match our career aspirations, whether we can run fast enough to still catch that bus, or whether we are capable of writing an eloquent introduction to a research article. In fact, we evaluate not only our own abilities but also the abilities of other people (Borkenau & Liebler, 1993; Fussell & Krauss, 1991). With respect to cognitive ability, we see differences in how smart our friends are, and we ponder how smart we are compared to them (Goethals & Klein, 2000). In order to arrive at such conclusions, we need to rely on information available from a variety of sources. As for the assessment of our own abilities, these sources comprise prior practical experience made with similar tasks, self-efficacy beliefs, level of aspiration, and even feedback in the form of results from standardized, objective, and usually scientifically devised psychometric ability tests (Ackerman & Wolman, 2007; Arsenian, 1942; Meyer, 1982). However, social comparison processes are the most important source. Such processes help us understand how our own abilities compare with those of other people by providing information, observed from a multitude of everyday situations, in relative terms (Guimond, 2006; Morse & Gergen, 1970; Mussweiler, 2003a, 2003b).

In general, the evaluation of such sources leads to self-estimates of abilities, such as cognitive ability, physical ability, social ability, and so forth, which are a vital part of a person's self-concept (Epstein, 1973). It can be argued that self-estimates of our own abilities should be relatively unbiased and closely related to our actual ability levels because we are constantly provided with real-life information about how we perform from all kinds of the abovementioned sources. However, many studies have found evidence that self-assessments are biased, mostly in the direction of a positively distorted self-evaluation (see, for instance, Brim, Glass, Neulinger, & Firestone, 1969; Gabriel, Critelli, & Ee, 1994; Kruger & Dunning, 1999; Maxwell & Lopus, 1994). According to this body of research, one well-documented distortion in self-assessment is the *better-than-average effect*, which is simply defined as a person's tendency to believe that her or his ability is above average (Guenther & Alicke, 2010). Such distortions are useful in helping individuals establish and maintain a positive self-concept because they facilitate self-esteem and feelings of self-worth and seem to be a necessary part of mental health (Taylor & Brown, 1988). Furthermore, less positively biased perceptions of the self and of the world are associated with depressive distortions and mental illness, although it is still up to discussion whether depression is associated with a realistic perception (de-

Philipp Alexander Freund, Department of Psychology, Leuphana University at Lüneburg, Lüneburg, Germany; Nadine Kasten, Department of Psychology, University of Osnabrück, Osnabrück, Germany.

Correspondence concerning this article should be addressed to Philipp Alexander Freund, Leuphana University at Lüneburg, Scharnhorststraße 1, 21335 Lüneburg, Germany. E-mail: afreund@leuphana.de

pressive realism; Alloy & Abramson, 1979, 1982) or an overly negative distortion (Beck, Rush, Shaw, & Emery, 1979; Dobson & Franche, 1989; Stone, Dodrill, & Johnson, 2001).

In this meta-analysis, the focus is on cognitive ability. Cognitive ability seems to be particularly important to people because it is intuitive to most that being smart has advantages, and we can see proof of that in many situations of everyday life. For example, people know that access to higher academic education is closely tied to performance at school and on standardized tests measuring scholastic ability (e.g., Freudenthaler, Spinath, & Neubauer, 2008). They also know that academic education facilitates the pursuit of financially lucrative careers, so that in essence, being smart pays off. Research provides further evidence that cognitive ability predicts a wide variety of life outcomes (e.g., Kuncel, Hezlett, & Ones, 2004). Among these outcomes are not only academic performance and occupational attainment but also many aspects of social life and, albeit indirectly, even mortality (Gottfredson, 1997b; Lubinski, 2000; O'Toole & Stankov, 1992; Schmidt & Hunter, 2004).

It is not just the actual cognitive ability level—which in many cases may in fact be unknown—that plays a role when individuals choose an education, career, and job. Self-estimations of cognitive ability are relevant as well because they reflect how people think about themselves in a subjective way, and they are also readily available to every individual. The misjudgment of abilities can generally have a considerable impact. For instance, overestimating one's car driving ability may lead to an accident. In this vein, the incorrect assessment of one's own cognitive ability almost invariably affects crucial life outcomes by decreasing the likelihood of achieving valuable goals (Ackerman & Wolman, 2007). Consider an individual's career choice: Overestimation of cognitive ability may lead to experience and feelings of failure and frustration, as the anticipated career may prove to be out of reach or too difficult to achieve success. On the flip side, underestimation of cognitive ability can result in boredom and, consequently, underperformance. For society as a whole, a close fit between ability and job demands is desirable because it helps decrease a waste of cognitive resources and expenditures on interventions (including occupational retraining and, in extreme cases, even psychotherapy) administered to individuals in the wrong occupational environments.

For cognitive ability, most empirical studies on the relationship between self-estimates and psychometric ability test scores report only weak to moderate correlations. These correlations typically range from about .19 to .39 (for a narrative review see Furnham, 2001). This suggests that while the relationship is significant, it appears to be relatively weak (based on the guidelines on the magnitude of effect sizes set by Cohen, 1988), implying that people are not very successful in estimating their own ability level.

One practical domain where self-assessments of cognitive abilities are routinely applied—despite or even because of such rather low correlations—is the field of vocational counseling, particularly in computer-assisted career guidance systems (e.g., Gati, Noa, & Krausz, 2001). Companies such as ACT, SHL, or Valpar International,[1] which are providers of career counseling advice that in part operate on a global scale, integrate explicit self-estimates of abilities into a complex electronic counseling process. The utility of self-estimates in career counseling is twofold. First, they can be related to the demand characteristics of occupational settings, so that counselees can compare their own self-assessed ability profile to occupational ability profiles. This helps them find occupations with a good fit to their own profile. Such an assessment therefore serves to identify similarities and dissimilarities between counselees' self-estimated ability profiles and the distinct ability profiles of occupations. It is especially suited to being implemented in electronic career guidance systems accessible via the Internet because it does not necessarily require the administration of a standardized, psychometric ability test as well. Instead, the focus is on the self-estimate as an expression of the counselee's self-concept with respect to her or his ability. Second, the degree to which self-estimated ability reflects actual ability as measured with a standardized ability test can be of interest to counselors. Constructivist career theories and qualitative career assessment methods (cf. McMahon & Patton, 2002; Young & Collin, 2004) call for and enable the counselor to not just apply diagnostic tests and give feedback on the results. Instead, and contrary to such a "test-and-tell" approach, the cognitions of the counselee are integrated into the assessment process in order to evaluate a counselee's self-concept and sense of identity. For instance, Hirschi and Läge (2008) showed how the use of self-estimates yields information beyond objective test scores in the domain of career interest assessment. Analogously, the subjective assessment of a counselee's cognitive ability can be compared to the results of a standardized ability test, and the degree of agreement between the two scores can be informative for the counselor and used in the counseling process because they indicate a counselee's degree of self-concept clarity and realism about personal aptitudes—both are desirable qualities for successful career decision making.

For both application purposes, the relationship between self-estimated and psychometrically assessed ability is critical to their success, albeit with a meaningful distinction. For the purpose of relating self-estimates to job demands, a small relationship simply induces substantial error into the analysis because there is no reliable information on the relationship between an individual's actual ability level and the demand characteristics of a specific occupation. Using self-estimates as proxies for "hard" ability would then be vague and imprecise. For the purpose of comparing self-estimated and psychometrically assessed ability level, it is this error (i.e., the difference between the two measures) that is of interest from the counselor's point of view. Therefore, even though the effectiveness of such career guidance systems is well documented (e.g., Gati et al., 2001; Sampson, 1994; Sampson & Watts, 1992), a crucial question is how well people can actually be expected to estimate their own level of ability. This concerns the aspect of the validity of self-estimated ability, where the relevant criterion is the score obtained from a psychometric ability test administered under objective, standardized conditions.

The goal of the present article is to investigate the validity of self-estimates of cognitive ability, as assessed through their relationship with psychometric ability test scores, by conducting a meta-analysis on the accumulated evidence in the field. Our article is structured as follows: First, we provide a theoretical and empirical framework in which self-estimates of cognitive ability and their relation to respective psychometric assessments using standardized tests can be integrated. This includes the definition of the

[1] See http://www.act.org, http://www.shl.com, and http://www.valparint.com, respectively.

essential concepts and related constructs, the introduction of extant theories, and the formulation of a general hypothesis. We also identify potential moderator variables that are expected to influence the validity of self-estimates and present corresponding hypotheses. Second, we detail the methods used in our meta-analysis, including our literature search strategy and study selection criteria, the coding process, and the applied analytic strategy. Third, we present the results for the overall relationship and for the moderator analysis. Fourth, the findings are critically discussed, implications for theory and practice are highlighted, limitations are considered, and venues for future research are identified.

Theoretical and Empirical Framework

Explicit and Implicit Theories of Cognitive Abilities

Human cognitive ability, often used synonymously with the term intelligence,[2] is one of the best researched, yet historically most controversially discussed, constructs in psychological research (Eysenck, 1998; Sternberg, 1985). In broad terms, cognitive ability is considered to be a very general category involving a wide variety of abilities, for instance, reasoning, problem solving, or abstract thinking (Gottfredson, 1997a). Historically, more specific attempts of theorists to converge on a uniform definition of cognitive ability have not been overly successful (Carroll, 1997; "Intelligence and Its Measurement: A Symposium," 1921; Sternberg & Detterman, 1986). Consequently, extant theories particularly differ in the number of factors considered to be indispensable for comprehending the nature of cognitive ability (Sternberg, 1985). In fact, the number of factors proposed in the major theories ranges from one (the so-called general factor of intelligence, often simply abbreviated as *g*; Spearman, 1927) to 150 or even 180 (Guilford, 1982, 1988).

However, psychological research in general rarely begins with fully agreed-upon definitions and consistent theories, though it may actually lead to them (Neisser et al., 1996). Thus, Carroll's (1993) stratum theory of cognitive abilities and its expansion to the Cattell–Horn–Carroll theory (CHC; Lohman, 2001; McGrew, 2005) can be viewed as a unifying framework, since it is "an expansion and extension of most of the previous theories of cognitive abilities" (Carroll, 2005, p. 74). Using factor analysis, Carroll (1993) reanalyzed more than 450 data sets and provided evidence that the abilities of interests could be clustered into three strata: Stratum III consists of one general factor, comparable to Spearman's (1927) *g*. At Stratum II, this general factor is split into several broad abilities, such as *gf* (fluid intelligence) and *gc* (crystallized intelligence), as hypothesized by Cattell (1971). Finally, at Stratum I, a large number of more narrowly defined abilities are located, for example, memory span or word fluency.

Despite all disputes on the nature of cognitive ability/intelligence in the research community, every individual has an idea of what constitutes an "intelligent" person. More precisely, besides experts—defined as people dealing intensively with the topic in either research or applied contexts—laypeople have a conception of what cognitive ability is, too, leading to so-called implicit theories of intelligence. These implicit theories have been objects of research for more than 60 years (e.g., Flugel, 1947; Furnham, 2001; Sternberg, Conway, Ketron, & Bernstein, 1981). Unlike explicit theories of intelligence, which are formulated and constructed by scientists, implicit theories do not need to be formally invented because they already exist (Sternberg, 1985). Another distinction is that implicit theories are also rarely systematically and empirically tested by their proponents (e.g., Ruzgis & Grigorenko, 1994).

With regard to the self-estimation of cognitive ability, such lay conceptualizations may directly affect expectations and evaluations of people's performance on cognitive ability tests (Sternberg, 2000). Thus, research on laypeople's perceptions of intelligence is of not only academic but also practical relevance (Furnham, 2001). There is substantial overlap between explicit and implicit theories. According to Sternberg (1985), this is hardly surprising because explicit theories are actually formalizations of experts' implicit theories. Thus, "traditional" facets of intelligence, such as verbal, numerical, or spatial ability, which can be located at the lower levels of the three-stratum CHC model, are frequently part of both explicit and implicit theories. However, some factors of intelligence mentioned in explicit theories are not shared by laypersons' perceptions and/or their relative importance is viewed differently (e.g., Sternberg et al., 1981). Here, cultural background is a factor to consider. For instance, in one study different ethnic groups have been found to feature different conceptualizations of "intelligence"—even though they all lived in North Carolina and were subjected to the same cultural mainstream (Heath, 1983). As another example, in Chinese culture, nonverbal reasoning has been found to be rated as more relevant to intelligence than verbal reasoning (Chen & Chen, 1988). Ultimately, intercultural differences in implicit theories can make it difficult to carry out meaningful comparisons of ability test scores across cultures (Greenfield, 1997). This is true from a conceptual as well as a methodological perspective (cf. Byrne et al., 2009). Concerning the latter, many studies neglect to pay sufficient attention to issues of measurement invariance across the groups under study. The comparison of test scores obtained from measures across groups requires that these scores possess the same psychometric properties in each group. The absence of measurement invariance is designated as measurement bias and—theoretically—prohibits using such test scores to make between-groups comparisons (cf. Meredith, 1993; Reise, Widaman, & Pugh, 1993; van de Vijver & Leung, 1997, 2000). Measurement invariance is usually investigated using methods of multigroup confirmatory factor analysis or item response theory (including methods for the analysis of differential item functioning, e.g., Vandenberg & Lance, 2000). It can be applied not only to scores on performance tests but also to all kinds of psychometric measures in general.

One construct that has repeatedly been discussed in the context of cultural influence on test performance is stereotype threat (Steele, 1997; Steele & Aronson, 1995). In general, stereotype threat implies that the performance on cognitive tasks of certain minority group members is affected by their anxiety that their performance may confirm a negative stereotype about their group. Wicherts, Dolan, and Hessen (2005) showed that issues of stereotype threat can be interpreted as a lack of measurement invariance and accordingly modeled in the same methodological framework.

[2] Throughout this article, when we use the two terms *cognitive ability* and *intelligence*, we refer to the same broad construct.

Self-Estimation and Self-Concept

Just like the definition of cognitive ability, the definition of self-estimation is rather imprecise and varies between studies. In order to come up with a working solution, we globally define self-estimation as a person's perception of her or his own abilities. Self-estimation is a process that is based on and involves repeated assessments in a variety of different concrete situations. Accordingly, this leads to domain-specific ability self-estimates. Self-estimation is different from the construct of self-concept, which is per se more general in scope (Marsh, 1990). More precisely, self-estimation is assumed to be an expression of self-concept, which by nature is "organized, multifaceted, hierarchical, stable, developmental, evaluative, and differentiable" (Shavelson, Hubner, & Stanton, 1976, p. 411). Following Epstein (1973), individuals use their self-concept as a self-theory, and self-estimations can be seen as (a) expressions of hypotheses to test this theory and (b) a means to assimilate new knowledge. As past experience with challenges posed by specific tasks in specific situations influences the formation of self-estimates of ability, such self-estimates can then be used individually as a basis for decision making and performance evaluation.

Although we constantly make self-estimates of our abilities, scientific research offers a rather unflattering picture regarding their degree of accuracy (cf. Ackerman, Beier, & Bowen, 2002). In order to explain why people make erroneous self-assessments, research has particularly focused on lack of metacognitive insight (Ehrlinger & Dunning, 2003). Empirically, Kruger and Dunning (1999) have provided evidence that people with low abilities and skills are especially affected by the tendency toward inflated self-views. In their study, the bottom quartile of performers in various ability domains (humor, logical reasoning, and English grammar) were most likely to overestimate their actual performance. Top performers, on the other hand, actually slightly underestimated their abilities. The extent to which this effect is attributable to statistical artifacts like regression to the mean (Nesselroade, Stigler, & Baltes, 1980) is still up to discussion (Ackerman et al., 2002; Krueger & Mueller, 2002). However, even after controlling for test score unreliability, the apparent asymmetry does not completely disappear (Kruger & Dunning, 2002).

Previous research has identified several variables and situational factors that influence the formation of distorted self-views (Alicke, 1985; Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Dunning, Meyerowitz, & Holzberg, 1989; Kruger, 1999). Probably the most influential of these is the specificity of the trait being evaluated: Given an ambiguous trait (such as the rather broad domain of general cognitive ability), an imprecise estimation of actual ability appears to be more likely (Ackerman et al., 2002). Dunning et al. (1989) suggested that content ambiguity contributes to idiosyncratic definitions of traits that allow for the maintenance of a positive self-view, which in general implies an overestimation of ability. A common theme expressed in the extant literature, therefore, is "that the ability to accurately judge one's own ability is sadly lacking" (Ackerman et al., 2002, p. 588).

Self-Estimated Versus Psychometrically Measured Cognitive Abilities

The framework in which the relationship between self-estimated and psychometrically measured cognitive abilities is embedded comprises theories taken from different areas of research in personality and social psychology. These theories share the assumption that people are motivated to evaluate their own abilities, but for quite different reasons. In this context, particular attention is given to the theory of self-consistency (Lecky, 1945), the theory of self-enhancement (Jussim, Yen, & Aiello, 1995; Swann, Griffin, Predmore, & Gaines, 1987), and Festinger's theory of social comparison processes (Festinger, 1954; Kruglanski & Mayseless, 1990). Self-consistency theory implies that people are motivated to form self-estimations that are consistent with their past experience in order to maintain a certain configuration of self-concept. Thus, prior experiences, but also constructs such as self-esteem, appear crucial for the prediction of the accuracy of self-estimated abilities. In contrast, self-enhancement theory proposes that people tend to view themselves as favorably as possible. For example, Paulhus, Lysy, and Yik (1998) showed that people rarely estimate their own skills to be below average (the better-than-average effect). Festinger's theory of social comparison processes emphasizes situational conditions, which can influence the accuracy of self-estimated intelligence scores. The primary assumption of this theory is that abilities can be compared to two main sources of information, namely, physical and social standards. Since physical standards are inaccessible when trying to self-estimate cognitive ability, ability estimates are primarily the result of comparisons with other people.

Empirically, the relationship between self-estimated cognitive abilities and psychometric test scores has been reported to be only weak to moderate. But despite such apparently rather discouraging findings, Holling and Preckel (2005) have argued that people are actually "more successful in estimating their general ability than the correlation between self-estimated and tested intelligence suggests" (p. 503). In an empirical study, Holling and Preckel reported a correlation of .46 (based on a sample of 88 participants) between self-estimates of intelligence and scores on an omnibus intelligence test (Intelligenzstrukturtest; IST 70; the IST 70 is a test constructed on the grounds of Thurstone's, 1938, theory of seven primary mental abilities; the IST 70 and its successor, the IST 2000-R, are among the most commonly applied omnibus intelligence tests in Germany). The standard deviation of the observed test scores was 9.58 (inferred from Holling & Preckel, 2005). Holling and Preckel then used this information to predict the psychometrically obtained test scores with the self-estimated scores by means of regression analysis. The standard error of estimation (SEE) in this analysis amounted to 8.51 points.[3] This suggests that approximately two thirds of the participants showed either a positive or a negative deviation of no more than 8.51 points from their tested scores. Using a correlation of .46 and a standard deviation of 15, which is the standard deviation on the IQ scale, leads to an SEE of 13.32 IQ points. These results are far from perfect, but they illustrate that medium-sized correlations can

[3] The formula for the SEE shows that mainly two factors influence its magnitude. These factors are the variance of the criterion scores (s_y^2) and the correlation between the dependent and independent variable(s): $SEE = \sqrt{s_y^2(1 - r^2)\frac{n-1}{n-2}}$. The ratio of $\frac{n-1}{n-2}$ is a correction factor that approaches unity for (very) large samples, and hence this factor does not substantially influence the SEE.

yield self-estimates that are arguably relatively useful because naturally, the scores obtained from standardized tests are not perfect measurements either, due to unreliability. Score reliability for intelligence tests is usually quite high, and an assumed reliability coefficient of, for example, .90 (which is considered a high level of reliability) implies a standard error of measurement (SEM) of 4.74. This shows that for the purpose of predicting test scores with self-estimates a correlation of .46 technically amounts to a degree of accuracy that is smaller by about factor 3 compared to the degree of accuracy that can be expected from the test score itself (conditional on a high psychometric test score reliability).

An interesting question is how accurate the prediction would be if the correlation between the two scores was higher, for instance, at $r = .80$. This would correspond to a large effect size according to the guidelines set by Cohen (1988). Assuming a correlation of .80 and a standard deviation of 15 leads to an SEE of 9.03 IQ points. This shows that even a large correlation does not lead to a much smaller SEE when the goal is prediction because the standard deviation in the criterion has such a large influence on the SEE. In fact, with a standard deviation as large as 15, one would need a correlation of .95 to obtain an SEE smaller than 5.00 (the SEE for a correlation of .95 is about 4.68). Extant empirical research has shown that the observation of such a close relationship between two psychological constructs is not very likely. In addition, it completely neglects the problem of measurement imprecision.[4]

Almost 30 years ago, Mabe and West (1982) conducted a meta-analysis to systematically investigate the relationship between self-estimated and psychometrically assessed ability measures. They reported an average correlation coefficient of .29. However, their meta-analysis did not focus just on cognitive ability. Instead, they included very different kinds of abilities (12 different abilities in total, among them also athletic, clerical, interpersonal, and managerial abilities and skills). With respect to the relationship between self-estimated intelligence scores and test scores, Mabe and West could rely only on 12 effect sizes. For this specific relationship, the average effect size was estimated at $r = .34$.

Mabe and West (1982) also analyzed the influence of a number of situational and experimental factors between studies on the effect sizes. According to them, particularly valid self-estimates can be expected if (a) a close fit between the self-assessed ability and the criterion measure is established; (b) the variable of interest is related to performance; (c) past test performance rather than future achievement is assessed; the estimate is made (d) as a social comparison (i.e., in relative terms) or (e) with reference to an explicit comparison group; (f) the corresponding ability distribution is characterized and described; (g) there is explicit assurance of anonymity; (h) the subjects have some experience in the self-evaluation of abilities; and (i) the subjects are expecting a comparison of their self-estimates to objective measures. Mabe and West labeled the presence of these factors *favorable measurement conditions*. The corresponding results in their meta-analysis, however, are for the complete set of 12 different abilities. Regarding the relationship between self-estimated and psychometrically assessed cognitive ability, Mabe and West were unable to conduct a moderator analysis and analyze the impact of the identified favorable measurement conditions due to the small number of effect sizes.

Many studies investigating the relationship between the two variables of interest report more than just one effect size. This leads to nonindependence of effect sizes because effect sizes from a single study share distinct characteristics (such as the same experimental setting, the same [sub]samples, etc.). Although in former meta-analyses researchers often chose to compute an average relationship for such studies, advancements in statistical methods (especially the advent of hierarchical linear modeling) now allow for the incorporation of multiple effect sizes per study by taking the cluster structure of effect sizes nested in studies into account. We detail this methodological approach in the Method section.

In the meta-analysis by Mabe and West (1982), the empirical basis for the correlation between self-assessed and psychometrically measured intelligence was very small, but it seems that this area of research has become a focal point of interest in the last 20 years, as indicated by the number of publications in the field: More than two thirds of the studies included in the present meta-analysis were published after 1990. In the same period, the availability of do-it-yourself tests has virtually exploded. There are various magazine and book publications on intelligence and intelligence tests; television shows featuring IQ tests are popular; and over the Internet, people can freely access all kinds of tests and questionnaires. Even though the scientific quality of many such tests may often be questionable, test takers have an opportunity to gain experience with different kinds of tasks, get feedback, and thus develop a better understanding of their own proficiency. An exemplary search using the keywords “free IQ test” in Google yielded over 3 million hits (search conducted on March 20, 2011). It could thus be argued that modern media, especially the Internet, have enormous potential to give laypeople a better understanding of a construct as elusive as cognitive ability, which may in turn lead to more valid self-estimates.

Hypotheses

Overall Relationship

Most studies investigating the relationship between self-estimates of cognitive ability and psychometric test scores report significant, positive correlations. In 1982, Mabe and West reported an average effect size of $r = .34$ (out of 12 effect sizes). We therefore expect to find a significant, positive overall relationship between the two variables.

Influence of Moderator Variables

We do not expect the average, true effect size to be the same for all cases in our analysis but instead assume significant variation. This is equal to assuming a random distribution of the true effect sizes. We further aim at explaining part of this variation through moderator variables. A total of six such moderators are included in this meta-analysis.

[4] Of course, it is not just the scores on the ability test that suffer from unreliability, as the self-estimate scores are affected by unreliability as well.

Methodology of self-assessment. Self-estimated cognitive ability scores can be assessed using different kinds of methodologies, the crucial point being whether the estimate is made with reference to a social comparison group or not. According to Festinger's theory of social comparison processes (Festinger, 1954; Kruglanski & Mayseless, 1990), it is plausible to assume that self-rating cognitive ability based on interindividual (social) comparisons leads to more valid estimates than estimates based on intraindividual comparisons. This is because intraindividual comparisons may prompt a person only to rank order their own abilities, without explicitly necessitating a comparison with others. Scales primarily evoking intraindividual comparisons feature labels with only absolute terms (such as *low ability* or *high ability*, or simply *bad* or *good*), but they do not mention a frame of social reference. Internal standards thus largely determine how a person will respond on such a scale. A person with high ability levels across all domains may still differentiate between these domains; for instance, she may evaluate her numerical skills to be better than her verbal skills, and therefore think that the latter are rather low compared to the former, while in fact, they may also be above average in comparison with the verbal skills of most other people. Also, a person with a "flat-plateaued" level of abilities may feel forced to differentiate between abilities that are equally high (or low), but without the comparison to other persons, the resulting self-estimated levels of ability are meaningless. A somewhat difficult aspect about the use of scales featuring absolute labels is that some people may still make interpersonal comparisons as well, which inserts another source of error.

Methods explicitly eliciting social comparisons are instead characterized by either the introduction and explanation of a normal distribution curve or by the use of a relative scale (i.e., the anchors of the scale are labeled in relative terms, such as ranging from *below average* to *average* to *above average*, including mention of a social comparison). Exact knowledge of and familiarity with the comparison group can provide a basis for even more valid self-evaluations (Martin, 2000; Mayseless & Kruglanski, 1987). For instance, senior high school students will have developed a good understanding of how capable they are in relation to their peers at their own school but may have difficulty assessing their ability when the reference group is "everybody in your age group" because a simple extrapolation is not possible. Reference to a specific comparison group should therefore add to the positive effect of relative scales. Recently, Goffin and Olson (2011) have discussed social–cognitive and evolutionary processes as reasons why ratings in relative terms can attain superior validity over absolute ratings and provided empirical illustrations using examples from three diverse research contexts (judgments in job performance measurement, the measurement of attitudes, and person perception).

A special case is the use of mixed scales, which are applied in various studies (see, e.g., Rammstedt & Rammsayer, 2002a, 2002b). While mixed scales feature anchors with labels in absolute terms, the middle category is labeled in relative terms (such as *average*). It is expected that mixed scales do not activate social comparison processes to the same degree as relative scales because they do not unequivocally address interindividual comparisons and may in fact even lead to confusion. Thus, with regard to the methodology of self-assessment, self-estimations made on relative terms are expected to be more valid than self-estimations made on absolute terms. The validity is expected to be even higher when a specific comparison group is mentioned. Mixed scale self-estimates should show no significant increase in validity over exclusively absolute self-estimates.

Ability type. The assessed ability varies substantively across studies, especially with respect to the degree of specificity. This also affects the choice of tasks used for the assessment. Omnibus tests of cognitive ability (viz., general intelligence) make use of a broad selection of tasks, so that strengths and weaknesses of individual test takers with regard to specific tasks are assumed to level out. The respective global scores therefore represent the top stratum of the CHC model. Many tests, however, use only a specific kind of task and accordingly assess a more narrowly defined ability, which can be located at a lower stratum in the CHC model. It seems reasonable to assume that the kind of measured ability affects the validity of self-estimates for two reasons. First, prior research has shown that a lack of familiarity with a task often leads to more erroneous self-estimates (Ng & Earl, 2008). Therefore, the more familiar people are with a task, the more valid the self-estimates should be. Since people in educational settings are usually confronted with tasks requiring verbal, numerical, and spatial abilities, they are familiar with and experienced in evaluating these, so that the degree of ambiguity should be rather small. Furthermore, these "traditional" abilities are emphasized in many common IQ tests and in popular books dealing with intelligence and its measurement (Furnham, 2000).

Second, tasks vary with regard to how salient they are for a test taker. Task salience is particularly tied to a layperson's conception of intelligence. Furnham (2001) suggested that most people consider numerical, spatial, and verbal abilities to be the "essence of intelligence" (p. 1401). Although other studies concerning laypeople's theories of intelligence focus on different abilities (e.g., Brim et al., 1969; Sternberg et al., 1981), the importance of verbal ability for implicit theories has been confirmed in most of them.

Third, it should be easier for individuals to compare their level of ability to that of others in the numerical, spatial, or verbal domains because opportunities for such comparisons are offered in many situations of everyday life. The abilities required in such concrete situations do not have to be integrated into a theoretically complete ability compound, as represented by general cognitive ability, and individuals are therefore assumed to be more successful at assessing them than general cognitive ability.

To sum up, we hypothesize that self-estimates concerning verbal, numerical, or spatial abilities should be more valid than self-assessments of general cognitive ability, which in turn is usually a compound of different subabilities (as implemented in omnibus test batteries). Consequently, use of these "standard" abilities should also result in more valid self-estimates than use of more rarely assessed abilities, such as memory or processing speed, for instance.

Order of assessment. Self-estimates require the availability of prior experience, which is usually evaluated in light of the current situation to which it is applied (Epstein, 1973). Here, proponents of self-consistency theory (Lecky, 1945) would argue that individuals strive to maintain a stable evaluation of their own abilities. However, the more the specific characteristics of a situation correspond to the conditions in which prior experience was made, the more valid self-estimates of intelligence should be. If there is a large difference between the tasks included in the psychometric measure and past experience, the specifics of the current situation could prompt test takers to either neglect their self-evaluations based on past experience or at least reassess them. Therefore, the order of assessment of self-estimated and psychometrically assessed intelligence can be expected to have an impact. If the psychometric test is taken first, the subsequent self-estimate should be influenced by this concrete (and vivid) experience, which in the usual study setup will have preceded the self-estimation by a comparably short time (of course, it is assumed that feedback of results will be postponed until after the collection of the self-estimates). Often, both assessments will even be tied together in applied contexts. In contrast, asking participants to give their self-estimates first means that they will need to rely exclusively on past experience that (a) may potentially differ quite a lot from the tasks used in the study, and/or (b) have a significant time lag, or (c) may even not be available to some participants. Accordingly, the assessment of future performance is likely to contain more error. We thus expect self-estimated cognitive ability to be more valid when self-estimates are made after the psychometric assessment. This is because self-consistency theory can in principle be used to account for both possible outcomes, but, as we argue, it is more appropriately tied to the scenario where the self-assessment is given first and there is no immediate task experience preceding the self-estimation process.

Gender of participants. Gender differences in self-estimates of cognitive ability have been and are still extensively discussed. There is ongoing discussion on whether differences in self-estimation between men and women are attributable to gender role expectations, or whether they can be ascribed to biological differences between the sexes (e.g., Greven, Harlaar, Kovas, Chamorro-Premuzic, & Plomin, 2009; Rosenberg & Simmons, 1975). We refer to theories highlighting the effects of socialization and gender stereotypes in particular, leading us to the term of *gender differences*. Many studies have found significant gender differences in the level of self-assessed intellectual abilities (e.g., Beloff, 1992; Byrd & Stacey, 1993; Hogan, 1978; von Stumm, Chamorro-Premuzic, & Furnham, 2009). Commonly, it is reported that women tend to underestimate, while men tend to overestimate, their own level of ability. Note, however, that this also depends on the task type at hand. Szymanowicz and Furnham (2011) provided meta-analytic evidence that there are moderate disparities between men's and women's self-estimations for general, mathematical, and spatial ability, but not for verbal ability.

Despite these findings, it is not really clear to what extent such gender differences lead to differential validity coefficients for self-estimates of ability. If, for instance, men tend to constantly overestimate their level of ability (which would mirror a greater tendency for men to employ self-enhancement strategies; Swann et al., 1987), this will not affect the magnitude of the relationship because it induces only an additive shift in the regression line (i.e., the results of the regression are invariant to linear transformations when standardized scores are used as predictor and criterion). Several studies indicate that men are more capable than women of giving valid self-assessments concerning their cognitive abilities, resulting in higher correlation coefficients (e.g., Furnham & Rawles, 1999). These findings are attributed to common gender stereotypes and other distortions, which lead to less valid estimations for females (Beloff, 1992). Thus, differences in math self-concept or mathematical self-evaluation (with mathematics perceived as a masculine domain, e.g., Bennett, 1997; Skaalvik & Skaalvik, 2004) are frequently consistent with traditional gender role stereotypes favoring boys (e.g., Byrne & Shavelson, 1986; Jackson, Hodge, & Ingram, 1994; Skaalvik & Skaalvik, 2004; Wilgenbusch & Merrell, 1999). This effect is often associated with gender stereotype threat. There is empirical support for a variety of variables influencing the degree of gender stereotype threat vulnerability (e.g., Brown & Josephs, 1999; Dar-Nimrod & Heine, 2006; Kiefer & Sekaquaptewa, 2007; Lesko & Corpus, 2006; Oswald & Harvey, 2000).

Just like stereotype threat with regard to belonging to a certain cultural group can be modeled and interpreted as a source of measurement bias (cf. Wicherts et al., 2005), it can be due to different expectations for female and male participants, which would explain the differential validity of self-estimates for the two gender groups. On the other hand, Reilly and Mulhern (1995) argued that the empirical evidence for gender differences in the validity of self-estimates appears to be a statistical artifact of noneliminated outliers; that is, only a small proportion of female respondents are responsible for these differences. Therefore, we do not expect substantial gender differences with regard to the validity of self-estimated intelligence.

Sample composition. Drawing samples from a restricted part of the population leads to range restriction and consequently to lower correlations (Alexander, 1988). Psychological studies often rely on student samples, which typically feature a large proportion of individuals with high cognitive abilities. Nonacademic samples in turn can be expected to show more variability concerning the key variables of interest. Therefore, correlation coefficients between self-estimated and psychometrically assessed cognitive ability obtained from nonacademic samples are expected to be higher than correlation coefficients obtained from purely academic samples.

Year of publication. The year of publication should have an effect on the correlation between self-estimated and psychometrically assessed cognitive ability because it can be assumed that people have a better understanding of the nature and demands of the respective tasks today than in the past. This enhanced understanding is mainly attributed to modern media—especially the Internet—providing easy access to the relevant information. We therefore expect more recent studies to report higher correlations than older studies.

The selection of moderators in this study does not completely cover the favorable measurement conditions identified in Mabe and West (1982). In particular, we were unable to find enough studies (or the relevant information was not given) on the role of anonymity versus nonanonymity in the experimental situation, prior experience with self-evaluation per se, and participants' expectations. These and other potentially relevant moderators (especially main effects of and interactions with culture) are therefore not investigated in the present analysis but are considered again in the discussion.

Method

Literature Search and Study Selection

We employed multiple search strategies in order to identify all relevant studies. First, we used the following keywords and their combinations for searches in the databases PsycINFO, ISI Web of Science, Google, and Google Scholar: *ability*, *cognitive*, *competenc**, *estimate**, *intelligen**, *perceive**, *self-apprais**, *self-assess**, *self-estimate**, *self-evaluat**, *self-perceive**, and *self-rate**. Second, we searched key journals, such as the *British Journal of Psychology*, *European Journal of Personality*, *Journal of Applied Psychology*, *Journal of Individual Differences*, *Journal of Personality*, *Journal of Personality and Social Psychology*, *Journal of Research in Personality*, *Intelligence*, and *Personality and Individual Differences*, for relevant articles. Third, we scanned the reference lists of all studies previously identified. Using these three strategies, we obtained a total of 238 studies with potentially relevant results.

For inclusion in the meta-analysis, a study had to meet the following standards: (a) The study had to compare a self-estimate of cognitive ability (intelligence) to a psychometrically assessed ability test measure. (b) This psychometric measure had to be collected by means of a standardized cognitive ability test. By standardized, we refer to tests that are administered in a standardized way, are objectively scored (either manually or electronically by the test administrator or a third person), and offer norms that allow inferences about the relative position of a test taker's score with regard to the norm sample. Such norms can either be included in the test manual or, mainly in studies where specifically developed test batteries are used, be derived from the data collected from the study sample. In order to ensure high levels of objectivity and comparability with regard to the criteria observed, other criterion measures, such as grades or experts'/others' (including teachers, parents, and peers) estimates of cognitive ability, were not considered. (c) A direct measure of self-assessed cognitive ability had to be applied. Studies with indirect measures of self-estimations (conclusions that were drawn from subjects' estimates of constructs that are related to cognitive ability, e.g., interest) were not included.

In total, 42 studies reported effect sizes that met these inclusion criteria. The primary reason for rejecting studies was the absence of a psychometrically sound test for the measurement of cognitive ability to which the self-estimate could be related. More than 70 studies (equal to about 29% of the studies identified in the search process) had to be dismissed for this reason.

Coding Process

We developed a standardized coding scheme based on our selection of moderator variables. The following characteristics were coded at the level of the individual effect size:

Methodology of self-assessment. Four categories were used (1 = *absolute scale*; 2 = *relative scale*; 3 = *relative scale including mention of a specific reference group*; 4 = *mixed scale*).

Ability type. Five categories were used (1 = *general cognitive ability*; 2 = *numerical ability*; 3 = *spatial ability*; 4 = *verbal ability*; 5 = *any other form of cognitive ability*).

Order of assessment. Three categories were used (1 = *self-estimation first*; 2 = *ability test first*; 3 = *unknown/cannot be derived from information in the study*).

Gender of participants. Three categories were used (1 = *mixed sample*; 2 = *female sample*; 3 = *male sample*).

Sample composition. Two categories were used (1 = *general sample*; 2 = *student sample*).

The following characteristic was coded at the level of the individual study:

Year of publication. The year of publication was registered as a continuous variable. All studies were mean-centered, with the year 1993 representing the mean.

Half of the studies (randomly selected) were first coded by the second author. They were then coded by another rater who was not familiar with the hypotheses. The average interrater reliability was satisfying at $\kappa = .87$. All discrepancies were discussed until consensus was reached. After this adjustment, the remaining studies were coded by the second author.

Computation of Effect Sizes

Since the included studies investigated the relationship between self-estimated and psychometrically assessed cognitive ability, most results were directly reported as correlation coefficients. In some cases, a t value for the significance test of r was reported instead of r. r was calculated from t using the standard computation formula (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009). If β coefficients from regression analysis were reported, the conversion formulae given in Peterson and Brown (2005) were used (this concerned two effect sizes). If estimating an effect size was not possible due to missing data, the respective case was excluded from the following analyses. This approach led to the exclusion of one study. In total, 154 effect size measures were obtained from the remaining 41 studies. Table 1 gives an overview of these studies including the specification of the corresponding effect size measures, the coded characteristics of the moderator variables, the number of items used for self-estimation, and the psychometric ability tests used. Most self-estimates (113, or 73.4%) were based on single-item measures; 17 self-estimates (11%) were based on multi-item measures; and for 24 self-estimates (15.6%), no information on the number of items was given.

In 33 cases (21.4%), a test battery was used. These test batteries were composed of different subtests that are usually selected on the grounds of arguments and requirements specific to the individual studies. Their use is predominantly prevalent in recent studies (especially in studies published by the workgroup of Ackerman and colleagues), but the oldest study in the data set, conducted by Cogan, Conklin, and Hollingworth (1915), also used a test battery. A number of tests (28, or 18.2%) rest on Thurstone's model of the seven primary mental abilities (these tests are the IST 70 and the IST 2000-R, the Leistungsprüfsystem, and the Wilde Intelligenz Test; all of these tests are in German). Furthermore, 22 (14.3%) effect sizes rest on the Differential Aptitude Test, 13 effect sizes (8.4%) rest on the Otis Quick-Scoring Test, and another 12 effect sizes (7.8%) rest on the Wonderlic Personnel Test. For the other 45 effect sizes (29.2%), a total of 17 different tests were used, among them the Baddeley Reasoning Test (6 effect sizes, 3.9%), the Wechsler Adult Intelligence Scales (4 effect sizes, 2.6%), Raven's Standard Progressive Matrices (4 effect sizes, 2.6%), or the SAT (3 effect sizes, 1.9%).

It is usually recommended to use the Fisher's z-transformed correlation coefficients in meta-analysis because their distribution is more normal than that of the Pearson correlation coefficients (Borenstein et al., 2009; Silver & Dunlap, 1987) and because the variance of the estimates of the correlation coefficients is not

Table 1
Overview of Included Studies and Effect Size Measures

Author	*r*	*z*	*n*	Students	Gender	Methodology of self-assessment	Ability type	Order of assessment	No. of SE items	Test
Ackerman et al. (2002)	−.05	−0.05	228	No	Mixed	Absolute	Verbal	Test	4	Test battery
	.47	0.51	228	No	Mixed	Absolute	Numerical	Test	3	Test battery
	.05	0.05	228	No	Mixed	Absolute	Other	Test	2	Test battery
	−.09	−0.09	228	No	Mixed	Absolute	Other	Test	2	Test battery
	.36	0.38	228	No	Mixed	Absolute	Verbal	Test	4	Test battery
	.18	0.18	228	No	Mixed	Absolute	Numerical	Test	3	Test battery
	−.09	−0.09	228	No	Mixed	Absolute	Other	Test	2	Test battery
	−.12	−0.12	228	No	Mixed	Absolute	Other	Test	2	Test battery
Ackerman et al. (2001)	.51	0.56	320	Yes	Mixed	Mixed scale	Verbal	Test	4	Test battery
	.40	0.42	320	Yes	Mixed	Mixed scale	Numerical	Test	5	Test battery
	.16	0.16	320	Yes	Mixed	Mixed scale	Other	Test	7	Test battery
Ackerman et al. (1995)	.29	0.30	93	Yes	Mixed	Mixed scale	Spatial	Test	N/A	Test battery
	.42	0.45	93	Yes	Mixed	Mixed scale	Verbal	Test	N/A	Test battery
	.58	0.66	93	Yes	Mixed	Mixed scale	Numerical	Test	N/A	Test battery
Ackerman & Wolman (2007)	.25	0.26	142	Yes	Mixed	Mixed scale	Verbal	Estimate	7	Test battery
	.48	0.52	142	Yes	Mixed	Mixed scale	Numerical	Estimate	8	Test battery
	.34	0.35	142	Yes	Mixed	Mixed scale	Spatial	Estimate	6	Test battery
	.27	0.28	142	Yes	Mixed	Mixed scale	General	Estimate	1	Test battery
	.25	0.26	142	Yes	Mixed	Mixed scale	Verbal	Test	7	Test battery
	.49	0.54	142	Yes	Mixed	Mixed scale	Numerical	Test	8	Test battery
	.39	0.41	142	Yes	Mixed	Mixed scale	Spatial	Test	6	Test battery
	.29	0.30	142	Yes	Mixed	Mixed scale	General	Test	1	Test battery
K. G. Bailey & Gibby (1971)	.10	0.10	112	No	Mixed	Relative	General	Test	1	Otis Quick-Scoring Test
	.46	0.50	124	No	Mixed	Relative	General	Estimate	1	Otis Quick-Scoring Test
K. G. Bailey & Lazar (1976)	.48	0.52	20	Yes	Female	Reference group	General	Estimate	1	Concept Mastery Test
	.53	0.59	20	Yes	Male	Reference group	General	Estimate	1	Concept Mastery Test
R. C. Bailey & Bailey (1974)	.40	0.42	37	No	Male	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.39	0.41	24	No	Female	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.29	0.30	40	No	Male	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.04	0.04	43	No	Female	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.55	0.62	44	No	Male	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.85	1.26	42	No	Female	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.27	0.28	45	Yes	Male	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
	.44	0.47	75	Yes	Female	Reference group	General	Estimate	N/A	Otis Quick-Scoring Test
R. C. Bailey & Mettetal (1977)	.49	0.54	139	No	Female	Relative	General	Estimate	1	Otis Quick-Scoring Test
	.35	0.37	139	No	Male	Relative	General	Estimate	1	Otis Quick-Scoring Test
Borkenau & Liebler (1993)	.32	0.33	100	Mixed	Mixed	Absolute	General	Estimate	1	Leistungsprüfsystem
Brim (1954)	.43	0.46	86	Yes	Mixed	Mixed scale	General	Test	1	American Council Psychological Examination
Chamorro-Premuzic et al. (2009)	.44	0.47	149	No	Mixed	Reference group	General	Unknown	1	Cognitive Ability Test
Chamorro-Premuzic & Furnham (2006)	.41	0.44	184	Yes	Mixed	Relative	General	Test	1	Wonderlic Personnel Test
	.41	0.44	184	Yes	Mixed	Relative	General	Estimate	1	Wonderlic Personnel Test
Chamorro-Premuzic et al. (2004)	.39	0.41	83	No	Mixed	Relative	General	Test	1	Wonderlic Personnel Test
	.49	0.54	83	No	Mixed	Relative	General	Test	1	Baddeley Reasoning Test
	.40	0.42	83	No	Mixed	Relative	Spatial	Test	1	S&M Spatial Ability Test
	.44	0.47	83	No	Mixed	Relative	Other	Test	1	Alice Heim Test

(*table continues*)

Table 1 (*continued*)

Author	*r*	*z*	*n*	Students	Gender	Methodology of self-assessment	Ability type	Order of assessment	No. of SE items	Test
Chamorro-Premuzic et al. (2005)	.22	0.22	182	Yes	Mixed	Relative	General	Estimate	1	Standard Progressive Matrices
Cogan et al. (1915)	.70	0.87	25	Yes	Male	Reference group	General	Estimate	1	Test battery
	.53	0.59	25	Yes	Male	Reference group	General	Estimate	1	Test battery
DeNisi & Shaw (1977)	.29	0.30	114	Yes	Mixed	Absolute	Verbal	Unknown	1	Personnel Test for Industry
	.36	0.38	114	Yes	Mixed	Absolute	Verbal	Unknown	1	SAT
	.41	0.44	114	Yes	Mixed	Absolute	Numerical	Unknown	1	Personnel Test for Industry
	.37	0.39	114	Yes	Mixed	Absolute	Numerical	Unknown	1	SAT
	.26	0.27	114	Yes	Mixed	Absolute	General	Unknown	1	Otis
	.35	0.37	114	Yes	Mixed	Absolute	General	Unknown	1	SAT
	.21	0.21	114	Yes	Mixed	Absolute	Spatial	Unknown	1	Minnesota Paper Forms Board Test
	.36	0.38	114	Yes	Mixed	Absolute	Other	Unknown	1	Bennett Mechanical Comprehension Test
Furnham (2005b)	.41	0.44	100	Yes	Mixed	Relative	General	Test	1	Wonderlic Personnel Test
Furnham (2009)	.44	0.47	187	Yes	Mixed	Relative	Verbal	Unknown	1	Multiple Intelligences Test
	.51	0.56	187	Yes	Mixed	Relative	Numerical	Unknown	1	Multiple Intelligences Test
	.37	0.39	187	Yes	Mixed	Relative	Spatial	Unknown	1	Multiple Intelligences Test
	.56	0.63	187	Yes	Mixed	Relative	Other	Unknown	1	Multiple Intelligences Test
	.38	0.40	187	Yes	Mixed	Relative	Other	Unknown	1	Multiple Intelligences Test
	.35	0.37	187	Yes	Mixed	Relative	Other	Unknown	1	Multiple Intelligences Test
	.31	0.32	187	Yes	Mixed	Relative	Other	Unknown	1	Multiple Intelligences Test
	.18	0.18	187	Yes	Mixed	Relative	Other	Unknown	1	Multiple Intelligences Test
Furnham & Chamorro-Premuzic (2004)	.30	0.31	187	Yes	Mixed	Relative	General	Estimate	1	Wonderlic Personnel Test
Furnham & Dissou (2007)	.53	0.59	101	Yes	Mixed	Relative	General	Test	1	Baddeley Reasoning Test
	.51	0.56	101	Yes	Mixed	Relative	General	Test	1	Baddeley Reasoning Test
Furnham & Fong (2000)	.19	0.19	172	Yes	Mixed	Relative	General	Test	1	Standard Progressive Matrices
Furnham et al. (2001)	.26	0.27	100	No	Mixed	Relative	Verbal	Estimate	1	Verbal Aptitude Test
	.35	0.37	100	No	Mixed	Relative	Numerical	Estimate	1	Numerical Aptitude Test
	.29	0.30	100	No	Mixed	Relative	Spatial	Estimate	1	Spatial Aptitude Test
Furnham et al. (2005), Study 1	.19	0.19	100	Yes	Mixed	Relative	General	Estimate	1	Baddeley Reasoning Test
	.27	0.28	100	Yes	Mixed	Relative	General	Estimate	1	Wonderlic Personnel Test
Furnham et al. (2005), Study 2	.27	0.28	131	Yes	Mixed	Relative	General	Estimate	1	Baddeley Reasoning Test
	.25	0.26	131	Yes	Mixed	Relative	General	Estimate	1	Wonderlic Personnel Test
Furnham & Rawles (1999)	.27	0.28	53	Yes	Male	Relative	General	Estimate	1	S&M Spatial Ability Test
	.09	0.09	140	Yes	Female	Relative	General	Estimate	1	S&M Spatial Ability Test
Furnham et al. (2006)	.29	0.30	64	Yes	Mixed	Relative	General	Test	1	Standard Progressive Matrices
	.47	0.51	64	Yes	Mixed	Relative	General	Test	1	Wonderlic Personnel Test
	.32	0.33	64	Yes	Mixed	Relative	General	Test	1	Baddeley Reasoning Test
Gabriel et al. (1994)	.27	0.28	62	Yes	Male	Relative	General	Estimate	1	Shipley Institute of Living Scale
	.30	0.31	84	Yes	Female	Relative	General	Estimate	1	Shipley Institute of Living Scale
Hodgson & Cramer (1977)	.71	0.89	34	No	Mixed	Absolute	Numerical	Test	N/A	Differential Aptitude Test
	.56	0.63	40	No	Female	Absolute	Numerical	Test	N/A	Differential Aptitude Test
	.15	0.15	34	No	Male	Absolute	Other	Test	N/A	Differential Aptitude Test
	−.25	−0.26	40	No	Female	Absolute	Other	Test	N/A	Differential Aptitude Test
	−.06	−0.06	34	No	Male	Absolute	Other	Test	N/A	Differential Aptitude Test
	.27	0.28	40	No	Female	Absolute	Other	Test	N/A	Differential Aptitude Test

Table 1 (*continued*)

Author	*r*	*z*	*n*	Students	Gender	Methodology of self-assessment	Ability type	Order of assessment	No. of SE items	Test
Holling & Preckel (2005)	.46	0.50	88	No	Mixed	Mixed	General	Unknown	N/A	IST 70
Kornilova et al. (2009)	.23	0.23	184	Yes	Mixed	Reference	General	Estimate	1	IST 70
Paulhus et al. (1998)	.20	0.20	174	Yes	Mixed	Absolute	General	Estimate	N/A	Wonderlic Personnel Test
	.24	0.24	174	Yes	Mixed	Absolute	General	Estimate	N/A	Wonderlic Personnel Test
	.23	0.23	241	Yes	Mixed	Absolute	General	Estimate	N/A	Wonderlic Personnel Test
	.26	0.27	241	Yes	Mixed	Absolute	General	Estimate	N/A	Wonderlic Personnel Test
Proyer & Ruch (2009)	.19	0.19	167	No	Mixed	Absolute	Verbal	Estimate	1	IST 2000-R
	.50	0.55	167	No	Mixed	Absolute	Numerical	Estimate	1	IST 2000-R
	.16	0.16	167	No	Mixed	Absolute	Spatial	Estimate	1	IST 2000-R
	.23	0.23	167	No	Mixed	Absolute	Other	Estimate	1	IST 2000-R
	.38	0.40	167	No	Mixed	Absolute	Other	Estimate	1	Standard Progressive Matrices
	.29	0.30	167	No	Mixed	Absolute	Verbal	Estimate	1	Wortschatztest
Rammstedt & Rammsayer (2002a)	.40	0.42	150	Yes	Mixed	Mixed	Verbal	Unknown	1	Leistungsprüfsystem
	.14	0.14	150	Yes	Mixed	Mixed	Verbal	Unknown	1	Leistungsprüfsystem
	.29	0.30	150	Yes	Mixed	Mixed	Numerical	Unknown	1	Wilde Intelligenz Test
	.27	0.28	150	Yes	Mixed	Mixed	Spatial	Unknown	1	Leistungsprüfsystem
	.17	0.17	150	Yes	Mixed	Mixed	Other	Unknown	1	Berliner Intelligenzstrukturtest
	.14	0.14	150	Yes	Mixed	Mixed	Other	Unknown	1	Leistungsprüfsystem
	.25	0.26	150	Yes	Mixed	Mixed	Other	Unknown	1	Leistungsprüfsystem
Rammstedt & Rammsayer (2002b)	.39	0.41	228	Yes	Mixed	Mixed	Verbal	Estimate	1	Leistungsprüfsystem
	.07	0.07	228	Yes	Mixed	Mixed	Verbal	Estimate	1	Leistungsprüfsystem
	.35	0.37	228	Yes	Mixed	Mixed	Numerical	Estimate	1	Leistungsprüfsystem
	.27	0.28	228	Yes	Mixed	Mixed	Spatial	Estimate	1	Leistungsprüfsystem
	.15	0.15	228	Yes	Mixed	Mixed	Other	Estimate	1	Leistungsprüfsystem
	−.04	−0.04	228	Yes	Mixed	Mixed	Other	Estimate	1	Leistungsprüfsystem
	.22	0.22	228	Yes	Mixed	Mixed	Other	Estimate	1	Leistungsprüfsystem
Reilly & Mulhern (1995)	.42	0.45	45	Yes	Male	Relative	General	Test	1	WAIS
	.15	0.15	80	Yes	Female	Relative	General	Test	1	WAIS
Steinmayr & Spinath (2009)	.15	0.15	136	No	Male	Mixed	Verbal	Estimate	1	IST 2000-R
	.46	0.50	136	No	Male	Mixed	Numerical	Estimate	1	IST 2000-R
	.33	0.34	136	No	Male	Mixed	Spatial	Estimate	1	IST 2000-R
	.23	0.23	136	No	Male	Mixed	Other	Estimate	1	IST 2000-R
	.21	0.21	203	No	Female	Mixed	Verbal	Estimate	1	IST 2000-R
	.40	0.42	203	No	Female	Mixed	Numerical	Estimate	1	IST 2000-R
	.30	0.31	203	No	Female	Mixed	Spatial	Estimate	1	IST 2000-R
	.09	0.09	203	No	Female	Mixed	Other	Estimate	1	IST 2000-R
Visser et al. (2008)	.31	0.32	200	Yes	Mixed	Reference group	Verbal	Unknown	1	Test battery
	.05	0.05	200	Yes	Mixed	Reference group	Spatial	Unknown	1	Test battery
	.38	0.40	200	Yes	Mixed	Reference group	Numerical	Unknown	1	Test battery
	.16	0.16	200	Yes	Mixed	Reference group	Other	Unknown	1	Test battery
	−.10	−0.10	200	Yes	Mixed	Reference group	Other	Unknown	1	Test battery
	−.01	−0.01	200	Yes	Mixed	Reference group	Other	Unknown	1	Test battery
	.20	0.20	200	Yes	Mixed	Reference group	Other	Unknown	1	Test battery
	.25	0.26	200	Yes	Mixed	Reference group	Other	Unknown	1	Test battery
	.20	0.20	200	Yes	Mixed	Reference group	General	Unknown	1	Test battery

(*table continues*)

Table 1 (*continued*)

Author	r	z	n	Students	Gender	Methodology of self-assessment	Ability type	Order of assessment	No. of SE items	Test
Webb (1955)	.21	0.21	95	No	Mixed	Relative	General	Estimate	1	Otis Quick-Scoring Test
Wells & Sweeney (1986)	.42	0.45	1,508	No	Mixed	Relative	General	Unknown	1	Quick test
Westbrook et al. (1994)	.66	0.79	111	No	Female	Reference group	Verbal	Test	1	Differential Aptitude Test
	.59	0.68	111	No	Female	Reference group	Numerical	Test	1	Differential Aptitude Test
	.49	0.54	111	No	Female	Reference group	Other	Test	1	Differential Aptitude Test
	.42	0.45	111	No	Female	Reference group	Other	Test	1	Differential Aptitude Test
	.47	0.51	111	No	Female	Reference group	Other	Test	1	Differential Aptitude Test
	.58	0.66	111	No	Female	Reference group	Spatial	Test	1	Differential Aptitude Test
	.57	0.65	111	No	Female	Reference group	Verbal	Test	1	Differential Aptitude Test
	.60	0.69	111	No	Female	Reference group	Verbal	Test	1	Differential Aptitude Test
	.62	0.73	99	No	Male	Reference group	Verbal	Test	1	Differential Aptitude Test
	.69	0.85	99	No	Male	Reference group	Numerical	Test	1	Differential Aptitude Test
	.47	0.51	99	No	Male	Reference group	Other	Test	1	Differential Aptitude Test
	.30	0.31	99	No	Male	Reference group	Other	Test	1	Differential Aptitude Test
	.35	0.37	99	No	Male	Reference group	Other	Test	1	Differential Aptitude Test
	.51	0.56	99	No	Male	Reference group	Spatial	Test	1	Differential Aptitude Test
	.69	0.85	99	No	Male	Reference group	Verbal	Test	1	Differential Aptitude Test
	.55	0.62	99	No	Male	Reference group	Verbal	Test	1	Differential Aptitude Test
Wolff & Wasden (1969)	−.25	−0.26	13	No	Female	Absolute	General	Estimate	N/A	WAIS
	.11	0.11	13	No	Female	Absolute	General	Test	N/A	WAIS

Note. For the moderator methodology of self-assessment: Absolute = absolute scale; relative = relative scale; reference group = relative scale including explicit mention of a reference group; mixed = mixed scale. For the moderator task type: General = general cognitive ability (intelligence); numerical = numerical ability; spatial = spatial ability; verbal = verbal ability; other = other ability. For the moderator order of assessment: Estimate = self-estimate first, psychometric test second; test = psychometric test first, self-estimate second. SE = self-assessment; IST 70 = Intelligenzstrukturtest; IST 2000-R = revised Intelligenzstrukturtest; WAIS = Wechsler Adult Intelligence Scales; N/A = not applicable.

independent of the population parameter, ρ.[5] Figure 1 shows that the 154 correlation coefficients appear to be relatively normally distributed. Using the nontransformed correlation coefficients offers the advantage that the results of subsequent moderator analyses can be directly interpreted in the original metric. We therefore decided to perform all analyses on the correlation coefficients. However, in order to check for the robustness of this decision, we also report the meta-analytically derived result for the overall relationship based on the Fisher's z-transformed coefficients.

Analytical Strategy

We conducted a random-effects meta-analysis where most studies in our data set report more than one effect size. In such cases, the main experimental settings are usually the same or at least very similar. In order to cope with any dependencies among these effect sizes, we used the hierarchical linear modeling (HLM) approach to meta-analysis (Raudenbush & Bryk, 2002). Here, the effect sizes are nested within the studies. For the included effect size $m = 1, \ldots, M_s$ in study $s = 1, \ldots, S$, the Level 1 model is simply a measurement model:

$$r_{ms} = \rho_{ms} + e_{ms}, \tag{1}$$

where r_{ms} is the sample estimate, ρ_{ms} is the corresponding population parameter, and e_{ms} is the sampling variance associated with r_{ms}, which is approximately normally distributed, $e_{ms} \sim \mathrm{N}(0,\upsilon^2)$. The variance υ^2 of the error term is given, which is why this is also called a *variance-known* model for meta-analysis (Raudenbush & Bryk, 2002). Since the 154 relationships are distributed over the 41 studies in our data set, we formulated both a within- and a between-study model (Bijmolt & Pieters, 2001; Kalaian & Raudenbush, 1996). At Level 2, the unknown true effect sizes depend on moderators coded at the relationship level (most of the moderators listed in the Coding Process section) and a random error, t_{ms}, which leads to the following within-study model:

$$\rho_{ms} = \beta_{0s} + \sum_{k=1}^{K} \beta_{ms} X_{kms} + t_{ms}, \tag{2}$$

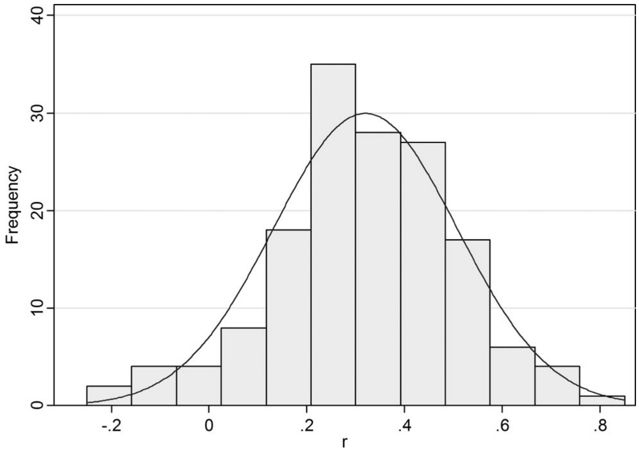

Figure 1. Histogram of correlation coefficients, with superimposed normal curve.

where β_{0s} is the intercept for study s; X_{kms} is the specific value on the experiment characteristic $k = 1, \ldots, K$; and β_{ms} is the corresponding coefficient. t_{ms} is assumed to be normally distributed, $t_{ms} \sim N(0,\tau^2)$. τ^2 is the variance of the true effect sizes.

At Level 3, the study-specific intercept β_{0s} is decomposed into moderators at the study level and another error term, u:

$$\beta_{0s} = \gamma_0 + \sum_{1=1}^{L} \gamma_l W_{ls} + u_s, \tag{3}$$

where γ_0 is the Level 3 intercept; W_{ls} is the specific value on a study characteristic $l = 1, \ldots, L$ (here, we use year of publication, centered at its grand mean); and γ_l is the corresponding coefficient. u_s is, again, assumed to be normally distributed, $u_s \sim \mathrm{N}(0,\varsigma^2)$, and ς^2 captures the variability in the true effect sizes between the studies. Technically, the unconditional, or "empty," model is a random-effects model, while the conditional model is a mixed-effects model because it includes fixed effects for the moderator variables in addition to the random components. One advantage of this approach is that the models for the calculation of the average effect size and for the moderator variables are nested and can directly be compared in terms of model fit through likelihood-ratio tests. All analyses were conducted with the software HLM 6.08 (Raudenbush, Bryk, Cheong, & Congdon, 2009).

Results

Overall Relationship

The 154 effect sizes ranged from $r = -.25$ to $r = .85$, with a mean of .32 and a standard deviation of .19. Over 93% (i.e., 144 of the 154) of the effect sizes were positive, indicating initial support for the first hypothesis, that the relationship between self-estimated and psychometrically assessed cognitive abilities is positive. The accumulated sample size for these 154 effect sizes was 22,256, which corresponds to an average sample size of about 145 participants per study (the minimum sample size was 13, and the maximum sample size was 1,508).

The result of the unconditional multilevel model, which is used to analyze the average effect size including error terms at the relationship and study level, closely resembled this finding. The intercept in this model was estimated at $r = .326$, with a standard error of 0.021 ($p < .001$). The 95% confidence interval for the mean effect size therefore ranged from 0.284 to 0.368. The variance component, τ^2, at the effect size level was estimated at 0.015, $\chi^2(113) = 517.381$, $p < .001$, and the variance component at the study level, ς^2, was estimated at 0.008, $\chi^2(40) = 118.843$, $p < .001$. Using the Fisher's r-to-z transformation and converting the estimate back to the original metric supports this estimate, as z was estimated at 0.337, which yields an r of .325. For z, $\tau^2 = 0.017$, $\chi^2(113) = 411.020$, $p < .001$; and $\varsigma^2 = 0.011$, $\chi^2(40) = 123.331$, $p < .001$. Together, these results suggest that adding moderator

[5] The variance of r is approximately $\frac{(1 - r^2)^2}{n - 1}$. For the Fisher's z-transformed correlation coefficients, the variance is approximately $\frac{1}{N - 3}$ (cf. Borenstein et al., 2009).

variables should help explain possible sources of variation among the effect sizes.

We also performed a funnel plot analysis to investigate if there was any kind of publication bias toward over- or underpowered studies in our data set. As Figure 2 shows, the majority of the effect sizes have rather low standard errors (smaller than 0.1), meaning that they were estimated with a reasonable degree of precision. Also, there appears to be a moderately asymmetric relationship between the magnitude of effect sizes and their standard errors, which becomes evident in the asymmetric distribution of effect sizes in the range where the standard errors are larger than 0.2 (cf. Egger, Davey Smith, Schneider, & Minder, 1997). Although there was one outlier (a correlation of .85 in a study by Bailey & Bailey, 1974, which corresponds to a Fisher's *z* coefficient of 1.26), this effect size did not have much leverage, and omitting it from the analysis had no impact on the stability of the results. Thus, there is practically no reason to assume that publication bias is a problem in the present meta-analysis. The funnel plot also gives evidence for heterogeneity among effect sizes.

Moderator Analysis

The model including the moderator variables is a straightforward extension of the basic, unconditional model for the overall effect. Before estimating this model, we checked for potential multicollinearity among the moderators at the experiment level. These moderators were all categorical, so we examined possible confounds among them using PRINCALS, a method of principal component analysis for categorical variables (cf. Bijmolt, van Heerde, & Pieters, 2005). According to Bijmolt et al. (2005), PRINCALS can be considered a conservative method for the detection of confounds among categorical moderator variables in meta-analysis. The threshold for confounds is typically set at $r = .5$ (cf. Bijmolt et al., 2005). For the moderator set in our analysis, we found a total of three correlations that exceeded an absolute value of .5 (–.626, –.629, –.640). The moderators concerned are participant gender and order of assessment (i.e., their dummy-coded categories). We consequently performed stability checks of our moderator analyses by omitting the variables related to these high correlations, one at a time. All these models yielded results very similar to those from the full model including all moderator variables. Hence, it can be concluded that multicollinearity was not a problem in our analysis, and we used all moderator variables in order to prevent any potential omitted variable bias.

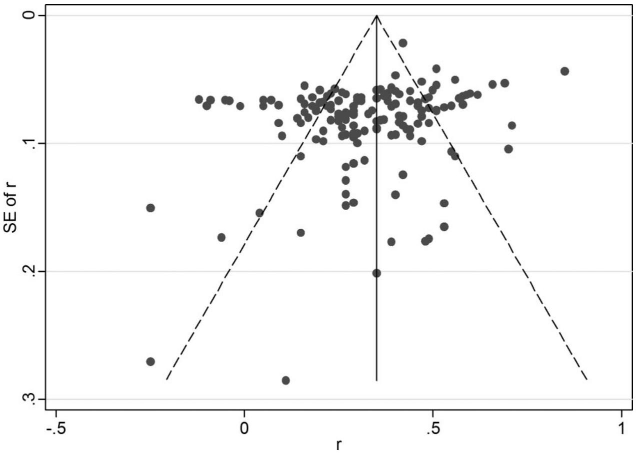

Figure 2. Funnel plot for the effect sizes.

Table 2 summarizes the results of the moderator analysis. The estimate for the intercept in this model is the configuration of all reference categories, that is, the average effect for a relationship observed with the use of absolute scales, assessing general cognitive ability, obtaining self-estimates first, having a mixed sample of nonstudent participants, and conduct of the study in the year 1993.[6] Below, we show how changing this configuration influences the size of the relationship.

Before testing the significance of the different categories of all the moderator variables featuring more than two categories (methodology of self-assessment [4 categories], ability type [5 categories], order of assessment [3 categories], and gender of participants [3 categories]) in comparison to their respective reference category, we performed separate omnibus tests of the relationship, which is conceptually similar to an analysis of variance approach. The corresponding test statistic is chi-square distributed, with degrees of freedom equal to $m - 1$, where m is the respective number of categories for the moderator variable.

For the moderator variable methodology of assessment, $\chi^2(3) = 7.534$, $p = .056$. For ability type, $\chi^2(4) = 65.811$, $p < .001$; for order of assessment, $\chi^2(2) = 2.093$, $p = .352$; and for gender of participants, $\chi^2(2) = 0.058$, $p > .500$. While these tests indicate that except for the moderator variable ability type, the differences between the categories of the moderators were not statistically significant, they are nondirectional, and they may mask potential single-category effects. Also, since we formulated directed hypotheses with regard to the effects of the different categories, we proceeded to test all single categories for their significance.

Methodology of self-assessment. Consistent with our hypothesis, and contrary to the suggestions of the omnibus test, there were significant differences between the four common methods of self-assessment. Estimates obtained from relative scales showed an improvement over estimates obtained from absolute scales (a regression coefficient, *b*, of 0.089, $p < .05$). Using an explicit reference group to which comparisons are to be made led to an even larger increase ($b = 0.136$, $p < .05$). In contrast, using a mixed scale, that is, a hybrid of a relative and an absolute scale, did not show a significant improvement over absolute scales ($b = 0.017$, $p = .706$).

Ability type. In line with our hypotheses, we found a large effect for the use of numerical abilities (vs. general cognitive ability), at $b = 0.161$ ($p < .01$). However, the effects for spatial and verbal abilities were both nonsignificant and thus suggest that assessing these abilities does not lead to more valid self-estimates ($b = -0.002$ and $b = 0.039$, respectively). As expected, assessing nonstandard cognitive abilities significantly decreased the size of the relationship ($b = -0.101$, $p < .01$).

Order of assessment. Contrary to our expectations, we did not find a significantly higher validity for self-estimates when the standardized test was applied first ($b = 0.045$, $p = .090$). For a relatively large number of effect sizes (35, or 22.7%), however, it

[6] The variable year of publication was centered around its grand mean, which in this case corresponds to the year 1993.

Table 2
Results of the Moderator Analysis

Variable	*m*	Coefficient	*SE*	95% CI	*p*
Fixed effect					
Intercept		0.261	0.041	[0.179, 0.343]	.000
Level 2 moderator variables					
Methodology of self-assessment[a]					
Relative scale	38	0.089	0.038	[0.013, 0.165]	.010[†]
Relative scale including reference group	43	0.136	0.064	[0.008, 0.264]	.017[†]
Mixed scale	46	0.017	0.045	[−0.073, 0.107]	.706
Ability type[b]					
Numerical ability	20	0.161	0.032	[0.097, 0.225]	.000[†]
Spatial ability	15	−0.002	0.035	[−0.072, 0.068]	.478[†]
Verbal ability	25	0.039	0.035	[−0.031, 0.109]	.132[†]
Other cognitive abilities	37	−0.101	0.036	[−0.173, −0.029]	.003[†]
Order of assessment[c]					
Standardized test first	57	0.045	0.033	[−0.021, 0.111]	.090[†]
Mixed/unknown	35	0.033	0.043	[−0.053, 0.119]	.221
Gender of participants[d]					
Females-only sample	26	−0.014	0.061	[−0.136, 0.108]	.816
Males-only sample	25	−0.010	0.058	[−0.126, 0.106]	.862
Sample composition[e]					
Students	86	−0.031	0.039	[−0.109, 0.047]	.220[†]
Level 3 moderator variable					
Mean year of publication = 1993		−0.002	0.001	[−0.004, 0.000]	.033[†§]

	Variance component	χ^2	*df*	*p*
Random effect				
Within study, τ^2	0.008	328.144	101	.000
Between studies, ς^2	0.006	114.120	39	.000

Note. Effects are reported in the Pearson *r* metric; sample size *m* is for number of effect sizes.
[a] Reference category: Absolute scales ($m = 27$). [b] Reference category: General cognitive ability ($m = 57$). [c] Reference category: Self-estimation first ($m = 62$). [d] Reference category: Mixed sample ($m = 103$). [e] Reference category: General sample ($m = 68$).
[†] *p* value halved due to one-tailed test. [§] Parameter becomes nonsignificant upon removal of two effect sizes (outliers with respect to year of publication; study by Cogan et al., 1915).

could not be determined whether the self-assessment or the standardized test was administered first, and the effect for this category (“mixed/unknown”) was not significant ($b = 0.033$, $p = .221$).

Gender of participants. These results were consistent with our hypothesis. Relying exclusively on either female or male samples did not significantly influence the validity of self-estimates of intelligence, compared to the standard case of using mixed samples ($b = –0.014$ and $b = –0.010$, respectively, both *p*s $> .500$).

Sample composition. The majority of the relationships (86 out of 154, or approximately 56%) were obtained from student samples, which we expected to lead to a lower validity due to range restrictions. In the present meta-analysis, we did not find a significant effect ($b = –0.031$, $p = .220$).

Year of publication. We found a significant effect for year of publication (mean-centered at 1993), at $b = –0.002$, $p = .033$. Note, however, that there was one study (Cogan et al., 1915) that represented an extreme outlier with regard to its year of publication. A subsequent analysis where this study and its two effect sizes were removed from the data set led to a change in the significance of the moderator ($b = –0.0008$, $p = .252$); the estimate for this parameter then became insignificant because the two effect sizes reported in Cogan et al. (1915) were both large and positive (*r*s = .70 and .53, respectively).

For the effect sizes included in our data set, the largest empirical Bayes estimate, at $r = .78$, was for a relationship obtained from a female sample of nonstudents, relative scales including explicit mention of a reference group, general mental ability, and administering the ability test first (the study was conducted in 1974). This correlation would lead to an SEE of 9.39 IQ points. In contrast, the SEE based on the overall estimate of $r = .33$ for the relationship between self-estimated and psychometrically assessed cognitive ability would be 14.16 IQ points, a difference of 4.77 IQ points.

In the moderator model, the variance component at the effect size level, τ^2, which is now the residual variance that is conditional on the moderator variables, was estimated at 0.008, $\chi^2(101) = 328.144$, $p < .001$, and the conditional variance component at the study level, ς^2, where year of publication is included as a moderator, was estimated at 0.006, $\chi^2(39) = 114.120$, $p < .001$. These coefficients show that significant variation in the true effect sizes can be explained by the Level 2 moderators, as is shown by a significant chi-square statistic for the comparison of the variance components between the unconditional and the conditional models, $\Delta\chi^2(12) = 189.237$, $p < .001$, and while at Level 3, year of publication significantly explained between-studies variation in the effect sizes, $\Delta\chi^2(1) = 4.723$, $p < .05$, this was no longer the case if we eliminated the two effect sizes from the Cogan et al.

(1915) study. Finally, we checked whether the conditional model did in fact exhibit superior fit over the unconditional model by conducting a likelihood-ratio test to assess the change in the deviance statistics between the models. For the unconditional model, $D = 663.974$, $df = 3$, and for the conditional model, $D = 600.920$, $df = 16$. The difference in the deviance statistics was 63.054, $df = 13$, which is also highly significant at $p < .001$, lending further proof that the inclusion of the moderators increases model fit.

Discussion

Interpretation of Results and Theoretical Implications

The two main goals of our meta-analysis were to (a) provide a synthesized mean effect size for the relationship between self-estimated and psychometrically assessed cognitive ability and (b) explain the variation between effect sizes as a function of moderator variables. These goals address two substantial questions: whether people are capable of estimating their own cognitive ability and which factors influence the validity of self-estimates of cognitive ability.

With regard to the first question, we found an average effect size of $r = .33$. Following Cohen (1988), this can be regarded as a medium-sized effect. It is also very similar to the mean effect size of $r = .34$ reported by Mabe and West (1982) almost 30 years ago, which was based on a small number of 12 studies only. Using this correlation to actually predict tested cognitive ability with self-estimates led to an SEE of 14.16 IQ points (assuming a standard deviation of 15, as in the standard-normal IQ distribution). This precision can be compared to the SEM of a standardized ability test with, for instance, a reported reliability coefficient of .90. The corresponding SEM is 4.74 IQ points, about one third of the self-estimate SEE of 14.16 IQ points. Thus, using self-estimates as proxies for standardized ability tests appears to be a rather imprecise endeavor, suggesting that the validity of self-estimates of cognitive ability is not very high. Note, however, that empirically obtained validity coefficients seldom surpass levels of $r = .50$—in fact, the well-acclaimed validity coefficient of general cognitive ability for the prediction of professional job achievement has been meta-analytically documented to be at $r = .51$ (corrected for range restriction and unreliability; Schmidt & Hunter, 1998).

With regard to the second question, however, the variance components at Levels 2 and 3 showed that there was significant heterogeneity among the effect sizes. This heterogeneity can be explained reasonably well through the moderator variables. Relative scales seem to prompt people to make interpersonal comparisons, which increase validity. They are even more valid if a specific reference group is named. Absolute scales, on the other hand, do not explicitly call for such comparisons. Based on our findings, the use of mixed scales cannot be recommended either, as they also seem not to directly facilitate comparisons to others. While the superiority of relative scales over absolute and mixed scales is hardly surprising, the degree of superiority was quite large. Apparently, although the accuracy of self-estimates primarily based on intraindividual comparisons (as initiated by the use of absolute scales) is not very high, utilizing social anchors to determine their own level of cognitive ability is of great help to individuals (see also Goffin & Olson, 2011). Considering social comparison theory (Festinger, 1954), these results show that explicitly making people self-assess their own cognitive ability in comparison to others significantly improves the magnitude of the relationship between self-estimated and psychometrically assessed cognitive ability. Asking for self-estimates, therefore, should always be done with relative scales and the mention of a specific reference group. However, it is conceivable that eliciting social comparisons may interact with phenomena such as stereotype threat. The activation of a negative stereotype for members of a minority group may not only impede their actual test performance (as has been shown extensively in the extant literature) but also insert a bias into the self-estimation process. Empirically, Aronson and Inzlicht (2004) provided evidence that stereotype threat vulnerability leads to inaccurate and unstable self-assessment of cognitive abilities. The authors suggested that the tendency to be affected by stereotype threat creates barriers to developing a stable and unbiased self-concept.

We expected numerical, spatial, and verbal abilities to show comparable levels of validity. Compared to the reference category (general cognitive ability), however, only self-estimates of numerical ability led to such an effect, while the validity of self-estimates of spatial and verbal abilities did not differ from that of self-estimates of general cognitive ability. Numerical ability seems to be especially salient and easy to self-estimate, whereas spatial and verbal abilities are not (cf. Ackerman & Wolman, 2007). One explanation for this could be based on how students make subject-specific experiences at school: Most people find out quite soon whether they are generally good or not at numbers and math, because the correct solutions, including the calculation method, are usually nonambiguous and give direct feedback of success. Because numerical ability tasks are similar to numerical tasks known from educational contexts, most people should have a good understanding of their numerical abilities. Verbal tasks, such as analogies, or odd-word-out tasks, are generally more open to interpretation, making it harder to clearly differentiate between correct and false answers. Also, these tasks differ hugely from spelling, grammar, and literature interpretation tasks typically encountered at school. While spatial abilities play a role at school (especially in science, e.g., three-dimensional representations of molecules etc.), concrete tasks concerning these abilities are quite often met in psychological tests for the first time, so that little to no experience is available (judging lengths and distances, a regular everyday task, is generally of only minor importance in tests of spatial ability). Previous results show a somewhat closer association of mathematical than of verbal self-concept or self-estimates with corresponding school achievement and thus support this assumption tentatively (Byrne & Gavin, 1996). However, these effects are rather inconsistent and need further empirical testing. Asking for self-estimates of less frequently assessed cognitive abilities nevertheless significantly decreases validity. These findings underline the importance of good familiarity, little ambiguity, and accumulated experience with the task type. Still, stereotype threat may again be a possible interacting factor here, with respect to the ability chosen for the self-assessment.

Regarding the order of assessment, we could not confirm our hypothesis that taking the standardized test first would lead to increased validity coefficients. This is a surprise insofar as the immediately preceding experience of working on tasks that represent the ability should facilitate the accuracy of self-estimates. The nonsignificance of this effect could be explained by self-consistency theory (Lecky, 1945). People have built an image of their abilities over a long time and will strive to prevent this image from being altered by a single, albeit recent, experience. Therefore, the validity of the ratings is not influenced significantly. In addition, feedback on the results from the standardized test is usually not given until after the self-estimates have been made, so people actually do not know how they or others fared on the test.

Most relationships were calculated for mixed samples. Relationships obtained from samples consisting exclusively of females or males, however, were neither significantly better nor worse. Apparently, there is no gender difference with respect to the validity of self-estimates. This finding does not support the assumption that female test takers experience feelings of stereotype threat that induce a kind of nonuniform measurement bias with regard to regression slopes for the relationship into the test scores (cf. Wicherts et al., 2005). There is a well-documented effect of gender differences with regard to the level of self-estimates of cognitive abilities (e.g., Furnham, Hosoe, & Tang, 2002; Rammstedt & Rammsayer, 2000; Szymanowicz & Furnham, 2011). However, this gender effect does not necessarily lead to decreased validity coefficients if it becomes equally effective for all subjects, because correlation coefficients are invariant to linear transformations when both the scores on the predictor and criterion are standardized. This corresponds to a measurement bias that affects only the between-groups intercepts. Furthermore, from a conceptual perspective, any kind of measurement bias with regard to participant gender would need to be explained in terms of how it is generated, developed, and (possibly) modified or sustained. Gender stereotypes are societal mechanisms that appear also to be influenced by culture (e.g., Guimond et al., 2007). While the results for the moderator variable gender of participants presented here seem to argue that there is no differential validity for females and males, they cannot go beyond this finding and explain the underlying mechanisms. In addition, it has to be noted that the majority of effect sizes were calculated on the basis of mixed samples consisting of both male and female subjects. More precisely, only 25 (or 16%) and 26 (17%) effect sizes were based on exclusively male or female samples, respectively. Thus, the absence of gender differences regarding the validity of self-estimates may also be due to lack of statistical power.

The nonsignificant effect for academic samples allows us to reject our assumption that range restriction leads to lower validity coefficients. There are at least two plausible explanations for this finding. First, academic samples in the field often consist of psychology students who gain a better understanding of cognitive ability than laypeople because of the subject they study. They also have more opportunities to collect experience with psychological tests and therefore know what to expect and how to react in such achievement situations. Second, a motivational bias that favors self-enhancement in job selection contexts is likely to be absent when the stakes are low. Such a detrimental bias can therefore arguably be avoided when researchers are relying on student samples.

With regard to the year of publication, we found no support for our hypothesis that more recent studies would lead to larger relationships after removal of the two effect sizes reported by Cogan et al. (1915), despite the widespread use and accessibility of ability tests through modern media, especially the Internet. This could mean three things. One, the year of publication of a study is not well-suited to function as a proxy for (real) experiences with ability tests; two, the availability of do-it-yourself tests does not correspond to making such experiences; or three, such experiences do not contribute to a more precise self-estimate. The results of the present analysis cannot provide a definite answer to this question, but we extend the discussion of this factor below.

Concluding from the results of the moderator analysis, the validity of self-estimates can be improved dependent on the configuration of the moderators, which in turn has an effect on the SEE. At $r = .78$, the highest empirical Bayes estimate implies an SEE of 9.39 IQ points. However, realizing the configuration associated with this estimate, which features a sample of female nonstudents, relative scales with explicit mention of a specific reference group, the assessment of general cognitive ability, and the administration of the ability test before the self-estimate is made, may not always be possible in practice. With respect to the factor methodology of self-assessment, it depends on what kind of reference group is available in a given situation. Quite often, there is no clearly defined reference group, or the reference group is very heterogeneous. It may in fact even be of interest to make people self-estimate their ability with reference to a heterogeneous reference group. Thus, implementation of this favorable measurement condition may be somewhat fuzzy. The choice of the ability type is also supposed to be dictated by the demands of the specific situation where the self-assessment is required. For example, it may simply be irrelevant to have an applicant for an interpreter job self-estimate her or his numerical ability, while the self-assessment of verbal ability would be of interest instead. From a substantive point of view, of the theories outlined in the background section, Festinger's (1954) theory of social comparison and Lecky's (1945) self-consistency theory receive support from the results of the present meta-analysis. However, it has to be noted that the theory of self-enhancement (Swann et al., 1987) cannot really be assessed with the data presented in the present meta-analysis, as the focus is solely on the validity of self-estimates, expressed through their correlations with psychometric test scores, and not on the degree of over- or underestimation of ability, information that would be needed in order to put self-enhancement theory to the test. In the next subsection, we discuss practical implications of the findings of the present meta-analysis.

Practical Implications

For practitioners, the rather low overall effect of $r = .33$ and the large influences of some of the moderator variables have a variety of implications. First, when counselees are asked to self-estimate their own ability, they should not be expected to be capable of giving a correct self-assessment, regardless of the concrete conditions under which the self-assessment is taking place. Instead, the differences between the self-estimate and the result from the standardized ability test can give meaningful diagnostic information about how realistically a person views her- or himself. This is underlined by the proportion of shared variance in the two constructs, which corresponds to just 10.89% for the overall relationship ($r = .33$), although it can be as high as 60.84% for the "optimal" measurement conditions identified by the moderator analysis ($r = .78$). This information is potentially relevant in career counseling (including counseling in school psychology) and personnel selection contexts in its own right. While the validity of psychometric ability tests (more precisely, general cognitive ability) for the prediction of academic and professional attainment has been shown to be superior to that of most other variables (at a corrected $r = .51$; cf. Schmidt & Hunter, 1998), there is little to no information on the (incremental) validity of self-estimated ability for the same objective criteria. It is conceivable that self-estimates of cognitive ability, representing core beliefs about one's own self-concept, influence the degree of effort people put forward when confronted with achievement tasks in educational and professional settings. This is a potentially useful piece of information that can be obtained easily. Of course, using self-estimates instead of psychometric ability tests for selecting job applicants is unrealistic because of self-enhancement bias (cf. Swann et al., 1987) that cannot be controlled (and neither could its degree be tested in the present meta-analysis). But their use appears to be promising in the job search domain (Prediger, 1999) and, with respect to job selection, could provide incremental validity when compared with the actual ability test scores. When self-estimates are used for this purpose, a careful and thorough instruction should also be included so that the reason for obtaining them becomes clear and individuals understand that they should attempt to be as accurate as possible in their self-assessment. Additional studies are needed here, as they would help advance the empirical state of research. Furthermore, Furnham (2005a) argued that self-estimates are relevant to job-related training and may even actually predict work performance beyond scores on psychometric ability tests in specific occupational contexts. On another note, self-estimates should also facilitate understanding of individual differences in career choice, decision making, and job performance (cf. Prediger, 1999; Wolman, 2009). However, all of these empirical considerations require at least reliable measurement of self-assessed cognitive ability. As can be seen in Table 1, the overwhelming majority of included effect size measures (a total of 113 correlation coefficients, or 73%) referred to self-estimates based on only one item. To ensure reliability and to increase validity more considerations are necessary regarding the measurement of self-estimated cognitive abilities. With respect to future studies, priority should be given to the development of appropriate psychometric measures of self-assessment.

Second, from a practitioner's point of view, the norms available for an ability test are crucial as well. Some test manuals offer a wide selection of norms for very different, and clearly defined, groups, while others provide rather global norms. Therefore, for some ability tests, using a specific reference group in order to obtain respective self-estimates is not feasible. In applied settings, it may also be an option to develop norms based on the data collected from prior and the present test administrations. The samples used for such norming processes can often be characterized as being rather specific. For instance, applicants for a university program in medicine may have to take a specific ability test, and if there is a large number of applicants per year, the development of norms for this kind of group becomes possible. Note, however, that in selection and placement settings, test security and test fairness are vital topics, so that test forms have to be equated because they are continually modified (e.g., Kolen & Brennan, 2004).

Third, the psychometric properties of test scores have to be taken into account by practitioners as well as by test takers. This begins with the degree of unreliability inherent in the scores, so that confidence intervals around test scores based on the test's SEM have to be computed. It further concerns the appropriateness of measurement models for the application to test data and also potential bias with regard to certain groups (as indicated by the absence of measurement invariance, respectively the presence of differential item functioning, cf. Reise et al., 1993; Vandenberg & Lance, 2000). Laypeople in general lack the knowledge to interpret ability test scores correctly and therefore need professional advice, lest they seriously misinterpret their results. Provision of this information is best practice.

Finally, accurate estimates of competencies and deficits are useful in educational settings highlighting self-assessments as a vital part of self-regulation (Eva & Regehr, 2007). Besides traditional models of teacher-directed learning, the concept of self-directed learning, which is associated with lifelong learning (Candy, 1991), has become a central issue in the study and practice of education (e.g., Boud, 1995; Garrison, 1997). Since the ability to direct and to regulate one's own learning is crucial to success in many professions and in other fields where knowledge is continuously evolving (Shokar, Shokar, Romero, & Bulik, 2002), curricula are designed to foster accurate self-assessment of competencies as a lifelong habit (Miflin, Campbell, & Price, 2000).

Critical Aspects of the Present Meta-Analysis

The results of our meta-analysis have to be interpreted cautiously with regard to the following factors.

Problems due to deficits in the empirical data basis. The empirical data basis does not allow for the investigation of other potentially relevant moderator variables. Mabe and West (1982) proposed nine measurement conditions influencing the validity of self-estimates. Using the studies included in our data set, we could not identify enough variation (if mentioned at all) with respect to the explicit assurance of anonymity, instructions emphasizing the comparison of self-assessments to objective tests, or prior self-evaluation experience, mainly because the number of studies reporting these experimental characteristics was insufficient. The same applies to moderator variables such as the influence of potential reward on the validity of self-assessment. This moderator variable is of particular interest because it seems reasonable that self-estimates made in selection settings should lead to less valid results simply because people are motivated to deliver as favorable an impression as possible. Virtually all effect sizes in the data set were obtained from low-stakes situations, where people are less likely to engage in impression management and individual outcomes are typically of only little importance. In low-stakes situations, individuals will further differ considerably with regard to test-taking (also labeled *current achievement*) motivation. Test-taking motivation has also been shown to be related to test performance (e.g., Freund & Holling, 2011; Freund, Kuhn, & Holling, 2011). When test takers are informed that their self-assessments are to be compared to their actual ability test scores, such a bias should be avoided (see also the subsection on practical implications above). Future studies should therefore explicitly investigate how the testing context influences the validity of self-estimates of cognitive ability.

Prior experience with cognitive ability tests, and with self-evaluation as well, can be strongly expected to facilitate the precision of self-estimates because naturally, experience entails performance feedback. Effects of feedback potentially have large implications (e.g., Berdie, 1954; Froehlich & Moser, 1954). Therefore, using the information on how one has performed on a test, or how precise one's self-estimation has been in the past, should influence the validity of self-estimates, as would also be predicted by self-consistency theory (Lecky, 1945). This point was also discussed by Mabe and West in 1982, but it appears that ensuing studies did not incorporate it to a sufficient degree. However, prior experience with ability tests entails another problem. There is ample meta-analytic evidence that retesting leads to higher test scores, and such score gains can even be increased with training (for a recent meta-analysis see Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007). Furthermore, the magnitude of these retest and training effects is affected by a number of moderators, among them the task type(s) used in the test. But while scores may be rising across test administrations, the actual ability level is assumed to be relatively stable. From a psychometric point of view, several researchers have shown that retesting induces measurement bias that makes it difficult to compare initial test and retest scores (e.g., Lievens, Reeve, & Heggestad, 2007). Individual test takers are therefore expected to make assessments that are affected by a number of psychometric problems, the concepts of which are difficult to understand for nonexperts.

These aspects also affect the implicit theories of test takers mentioned in the background section. A personality construct relevant to implicit theories is the incrementalist/entitist personality type as introduced by Dweck (e.g., Dweck, 2006; Dweck & Leggett, 1988). According to Dweck, individuals' implicit theories about abilities differ with regard to how malleable they think such abilities are. An individual who believes that abilities can be improved is called an *incrementalist* with a "growth" mindset, while an individual who believes that abilities are stable and cannot really be improved is an *entitist* with a "fixed" mindset. Meta-analytic results on how retesting and training help to increase scores on cognitive ability tests (e.g., Hausknecht et al., 2007; Kulik, Bangert-Drowns, & Kulik, 1984; Kulik, Kulik, & Bangert, 1984) lend support to the appropriateness of a growth mindset, which in turn can be assumed to lead individuals to reflect on their performance more deeply and critically assess their errors in order to improve. This may subsequently affect the processes involved when people are asked to self-estimate their own cognitive ability.

Valid self-evaluation is also critical for self-organized learning across the life span. Experience and training can help enhancing agreement of self-evaluation with objective criteria (Sluijsmans, Dochy, & Moerkerke, 1999). Again, Dweck's incrementalist/entitist personality type could have an impact here. People who are incrementalists should be expected to more strongly engage in metacognitive thinking about their own performance on cognitive tasks and strive to get better. In order to do so, they invariably develop a better understanding of where different performance levels rank and accordingly become better at comparing their own ability to that of others. In contrast, entitists tend to believe that their ability levels are fixed and cannot really be improved, so they may engage less in thinking about what different ability levels really mean and, especially, how relative they actually are. In the studies included in the present meta-analysis, information about whether individual participants had prior experience either with ability tests or with self-estimation of ability could not be obtained, and neither could it be determined if the participants in a particular study were incrementalists or entitists, but it appears promising to conduct studies investigating these factors.

Based on the findings of Kruger and Dunning (1999), who reported that mainly individuals in the lower area of the ability distribution tend to heavily misjudge their ability level, it would be interesting to investigate if actual ability level has an influence on the magnitude of the relationship between self-estimates and psychometric test scores. However, the primary studies included in the present meta-analysis generally do not provide the necessary information to do so. One would at least need separate effect sizes for distinct ability groups. This aspect is therefore better suited for investigation in a primary study.

The vast majority of the studies were conducted in Western countries, using student samples. There is ongoing discussion on the generalizability of such results. Henrich, Heine, and Norenzayan (2010) dubbed the common study sample as WEIRD (participants from Western, Educated, Industrialized, Rich, and Democratic societies). Cultural differences can be expected to play a huge role, in particular regarding conceptions of intelligence (Chen & Chen, 1988; Heath, 1983; Nisbett, 2003; Sternberg & Kaufman, 1998), test performance (through stereotype threat; e.g., Steele, 1997; Steele & Aronson, 1995), and the importance of self-impression management (Heine, Lehman, Markus, & Kitayama, 1999). In many Eastern (i.e., primarily Asian) countries, for instance, social interaction abilities are considered to be more important than individual cognitive abilities (e.g., Nevo & Khader, 1995). Accordingly, these factors could affect the magnitude of the relationship between self-estimates and tested cognitive ability. Because the results obtained from psychometric tests are used as criteria for advancement in educational and professional careers, such cultural differences in how the psychological constructs being measured are viewed and interpreted can also be expected to exert a direct influence on life outcomes. If, for instance, a member of a minority group applies for a job and is asked to take a test that has been found to function well with members of the majority group, it needs to be ascertained that the test's applicability can be generalized across groups. It is considered best practice to make sure that no individual, regardless of her or his cultural or societal background, has any kind of disadvantage due to biased tests (Byrne et al., 2009). Psychometric measures of any kind used in such contexts (including intelligence tests) should therefore routinely be checked, validated, and, if necessary, modified. This should ultimately lead to improved quality—and acceptance across groups—of tests. Additionally, gender differences in self-estimates of IQ do in part depend on cultural factors (Furnham, 2001), casting doubts on the cross-cultural generalizability of our meta-analytical results. Gender issues, of course, can be assessed and modeled in the same methodological framework as any other kind of between-groups issue.

As a starting point for investigating cultural impact on the meaning of psychological constructs such as cognitive ability (in performance and self-estimation), comparisons between traditional "Western" and "Eastern" samples appear interesting and practically feasible, but they should not stop there. In the data set of the present meta-analysis, there are not enough studies investigating the relationship between self-estimated and psychometrically assessed cognitive ability in Eastern samples, so a comparison between effects found in Western and Eastern samples could not be carried out.[7]

Cross-cultural designs are generally desirable in modern empirical studies investigating relationships between psychological constructs (e.g., Segall, Lonner, & Berry, 1998) because they help overcome problems associated with WEIRD samples. They also appear to be particularly promising for a topic like that of the present meta-analysis, as they can address similarities and differences between cultures with regard to measures of self-assessment and psychometric measures by analyzing questions of measurement invariance (e.g., van de Vijver & Leung, 1997, 2000) and how these two kinds of measures are interrelated. As discussed above, it is possible, for instance, that biasing phenomena such as stereotype threat could be triggered for members of different cultural groups when tests ask for social comparisons in the self-estimation process. These arguments help to show that investigations into the psychometric properties of test scores are not a purely technical endeavor but instead contribute to the sound methodological conduct of empirical studies. They can readily be carried out if researchers use adequate study designs. In order to arrive at a genuine and deep understanding of how psychological processes function in individuals, there is a definitive need to carry out studies under a much wider variety of conditions, in order to overcome the many problems associated with WEIRD research. The subject of the present meta-analysis integrates psychological constructs from the cognitive domain that play a huge role in everyday life, and it is therefore desirable that the best available research methods should be applied to advance the field. This is clearly a goal future studies should strive to attain.

Furthermore, although in the present meta-analysis we have examined a number of moderators at the experiment level (i.e., situational and methodological moderators), we were unable to investigate moderators influencing the genesis and stability of self-estimates. There are at least three factors that are of potential interest here. The first factor is personality (especially the Big Five), which is assumed to have a substantial impact on self-estimates (Chamorro-Premuzic & Furnham, 2004). Also, constructs with a close proximity to task interest and performance, such as need for cognition (Cacioppo & Petty, 1982) or typical intellectual engagement (Goff & Ackerman, 1992), appear promising. The second factor is perceived task difficulty. Lichtenstein and Fischhoff (1977, 1980) were able to show that easy tasks lead to underestimates of ability, while hard tasks lead to overestimates. Finally, self-worth may play a significant role in the development of self-estimates, as self-enhancement bias can help to protect against detrimental, negative self-estimates in light of poor performance (Trope, 1986).

Corrections for Unreliability and Other Artifacts

Correcting for study artifacts is common practice in meta-analyses aiming at validity generalization in order to assess the magnitude of the relationship at the construct level. In particular, the imperfect reliability of both the predictor and the criterion impose a lower bound on the observed relationship (Hunter & Schmidt, 2004). Reliability, of course, is always a property of test scores and never of instruments (e.g., Nunnally & Bernstein, 1994). Therefore, it is necessary that information on the reliability of test scores in primary studies is communicated. While providing information on test score reliability has become standard practice for most scientific journals, this information is often given rather globally. For instance, if some kind of test is applied, but the results are reported at the level of subgroups, information on test score reliability is often restricted to the level of the entire sample, and it would be incorrect to assume the same level of reliability for the subgroup scores. In the data set used for the present meta-analysis, we found usable information on the reliability of the test scores for only 14 of the 154 effect sizes. This is a percentage of just about 9%.

Usable information on the reliability of self-estimates was provided more often, for 54, or about 35%, of the 154 effect sizes (as a trend, studies reporting the appropriate information on reliabilities of self-estimates and test scores were among the more recent studies). Here, the unweighted average reliability coefficient (representing different kinds of reliability coefficients, but primarily estimates for Cronbach's alpha) was .77, with a standard deviation of .14. The range of these coefficients was between .32 and .97. This indicates a wide range of reliability coefficients, and in some studies, the self-estimate scores clearly suffered from a high degree of unreliability, which may be due to choice of method, number of items (approximately 73% of the self-estimates were based on only one item), or other potential sources. However, both figures are far from adequate to correct for the unreliability in either the criterion (standardized test scores) or the predictor (self-estimates of ability). It is therefore recommended that future studies report the necessary information at the respective levels and not just provide information on global test score reliability.

Number of Items Used for Obtaining Self-Estimates

As indicated above, most effect sizes in the present meta-analysis were based on self-estimates obtained with one-item measures. In contrast, the ability measures were (usually) aggregate scores based on a much larger number of items. Besides expectable differences in the reliability of these scores, this corresponds to an asymmetrical relationship because an aggregate score covers a variety of different sources and is therefore much more diverse than a one-item measure, which explicitly demands cognitive integration of a number of concrete experiences (in this case, attempts to solve a number of items on an ability test). Hence, balancing out the level of aggregation could increase the validity of self-estimates. To our knowledge, empirical studies investigating such effects are lacking, so this is another topic future studies should look into.

Conclusions

With regard to the meta-analysis by Mabe and West (1982), which was conducted almost 30 years ago, it is interesting that the overall relationship remains virtually unchanged (meta-analytically derived effect sizes of $r = .34$ vs. $r = .33$). On the one hand, this indicates temporal stability of the relationship. It seems

[7] On a side note, we have to acknowledge that our literature search findings were restricted to work reported in either English or German. Potentially relevant work conducted and reported in other languages (e.g., Chinese, French, Russian, or Spanish, etc.) was not included in the present meta-analysis.

that while people are somewhat capable of giving valid self-estimates, especially under favorable conditions, the degree of accuracy is genuinely limited. On the other hand, a critical assessment of the studies carried out post-Mabe and West shows that the majority of them did not incorporate all of the—and often not even the most influential—favorable measurement conditions suggested in order to maximize the relationship between self-assessed and psychometrically measured (cognitive) ability. In fact, the majority of effect size measures published after 1982 adhere to only three favorable measurement conditions or fewer (79 of 120, or 66%). This can be interpreted in at least two ways. One, the suggested measurement conditions are not easy to establish. This excuse may be deemed valid only for Mabe and West's measurement condition (h), which states that subjects should have experience in the self-evaluation of abilities. Clearly, this condition is unrealistic to uphold in a typical empirical study. All the other favorable measurement conditions appear readily implementable, conditional on whether researchers strive to pursue finding the maximum relationship between the two constructs. Two, studies conducted since the publication of Mabe and West's 1982 article deliberately vary with regard to adhering to these favorable measurement conditions in order to test their impact empirically. This would result in stringent tests of Mabe and West's advice and consequently an advance in the field. Upon closer inspection, it seems that a sizable number of studies were not specifically designed to put the favorable measurement conditions suggested by Mabe and West to the test. This may be at least in part be due to the primary focus of these studies, which was not always the relationship of interest to the present meta-analysis.

As Henrich et al. (2010) pointed out, the generalizability of results from studies investigating psychological phenomena is limited when these studies do not cross cultural boundaries. We agree with their recommendations to conduct research beyond a WEIRD context. In this regard, the use of suitable statistical methods is also recommended, for instance, in the analysis of potentially debilitating testing effects, such as stereotype threat (cf. Wicherts et al., 2005). These methods help to answer questions such as if observed differences between groups (cultural, gender, age, etc.) are true differences or due to measurement bias. For the present meta-analysis, we have to confine our interpretations to the current (up to 2011) empirical database and cannot address the kind of measurement issues discussed in this article.

To conclude, based upon the results of the present meta-analysis, particularly valid self-estimates can be expected if individuals are requested to provide relative appraisals of their numerical ability compared to a clearly specified comparison group.

Our meta-analysis offers new insight into the relationship between self-estimated and psychometrically assessed cognitive ability. It illustrates the variability among effect sizes and shows under which conditions more valid self-estimates can be obtained. It appears that self-estimates provide interesting diagnostic information in their own right, but it remains to be shown how useful this information can really be in applied contexts. However, many potential moderators could not be investigated because they have not been implemented in empirical studies. Future studies should therefore explicitly integrate such promising moderators in order to provide further information on how the validity of self-estimates of cognitive ability can be enhanced. We have argued in different places that the relationship between self-assessed and psychometrically measured cognitive ability may be affected by biasing phenomena, especially stereotype threat, which can manifest itself in various ways. In multigroup as well as longitudinal study designs, there is a definitive need for researchers to investigate issues of measurement invariance. In such cases, if measurement invariance is not tenable, the interpretation of score differences across groups and/or time becomes difficult. Researchers should therefore also take advantage of modern psychometric and statistical methods, such as latent variable, multigroup, and multilevel modeling.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33,* 587–605. doi:10.1016/S0191-8869(01)00174-X

*Ackerman, P. L., Bowen, K. R., Beier, M. E., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology, 93,* 797–825. doi:10.1037/0022-0663.93.4.797

*Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and no cognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied, 1,* 270–304. doi:10.1037/1076-898X.1.4.270

*Ackerman, P. L., & Wolman, S. D. (2007). Determinants of validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied, 13,* 57–78. doi:10.1037/1076-898X.13.2.57

Alexander, R. A. (1988). Group homogeneity, range restrictions and range enhancement effects on correlations. *Personnel Psychology, 41,* 773–777. doi:10.1111/j.1744-6570.1988.tb00653.x

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49,* 1621–1630. doi:10.1037/0022-3514.49.6.1621

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average-effect. *Journal of Personality and Social Psychology, 68,* 804–825. doi:10.1037/0022-3514.68.5.804

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441–485. doi:10.1037/0096-3445.108.4.441

Alloy, L. B., & Abramson, L. Y. (1982). Learned helplessness, depression and the illusion of control. *Journal of Personality and Social Psychology, 42,* 1114–1126. doi:10.1037/0022-3514.42.6.1114

Aronson, J., & Inzlicht, M. (2004). The ups and downs of attributional ambiguity: Stereotype vulnerability and the academic self-knowledge of African American college students. *Psychological Science, 15,* 829–836. doi:10.1111/j.0956-7976.2004.00763.x

Arsenian, S. (1942). Own estimate and objective measurement. *Journal of Educational Psychology, 33,* 291–302. doi:10.1037/h0063257

*Bailey, K. G., & Gibby, R. G. (1971). Development differences in self-ratings on intelligence. *Journal of Clinical Psychology, 27,* 51–54. doi:10.1002/1097-4679(197101)27:1<51::AID-JCLP2270270108>3.0.CO;2-2

*Bailey, K. G., & Lazar, J. (1976). Accuracy of self-ratings of intelligence as a function of sex and level of ability in college students. *Journal of Genetic Psychology, 129,* 279–290.

*Bailey, R. C., & Bailey, K. G. (1974). Self-perceptions of scholastic ability at four grade levels. *Journal of Genetic Psychology, 124,* 197–212.

*Bailey, R. C., & Mettetal, G. W. (1977). Perceived intelligence in married partners. *Social Behavior and Personality, 5,* 137–141. doi:10.2224/sbp.1977.5.1.137

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York, NY: Guilford Press.

Beloff, H. (1992). Mother, father and me: Our IQ. *The Psychologist, 5,* 309–311.

Bennett, M. (1997). Self-estimates of ability in men and women. *Journal of Social Psychology, 137,* 540–541. doi:10.1080/00224549709595475

Berdie, R. F. (1954). Changes in self-ratings as a method of evaluating counseling. *Journal of Counseling Psychology, 1,* 49–54. doi:10.1037/h0056290

Bijmolt, T. H. A., & Pieters, R. G. M. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters, 12,* 157–169. doi:10.1023/A:1011117103381

Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research, 42,* 141–156. doi:10.1509/jmkr.42.2.141.62296

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, England: Wiley. doi:10.1002/9780470743386

*Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65,* 546–553. doi:10.1037/0022-3514.65.3.546

Boud, D. (1995). *Enhancing learning through self-assessment.* New York, NY: Routledge Falmer.

*Brim, O. G., Jr. (1954). College grades and self-estimates of intelligence. *Journal of Educational Psychology, 45,* 477–484. doi:10.1037/h0057492

Brim, O. G., Jr., Glass, D. C., Neulinger, J., & Firestone, I. J. (1969). *American beliefs and attitudes about intelligence.* New York, NY: Russell Sage Foundation.

Brown, P. B., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76,* 246–257. doi:10.1037/0022-3514.76.2.246

Byrd, M., & Stacey, B. (1993). Bias in IQ perception. *The Psychologist, 6,* 16.

Byrne, B. M., & Gavin, D. A. (1996). The Shavelson model revisited: Testing for the structure of academic self-concept across pre-, early, and late adolescents. *Journal of Educational Psychology, 88,* 215–228. doi:10.1037/0022-0663.88.2.215

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3,* 94–105. doi:10.1037/a0014516

Byrne, B. M., & Shavelson, R. J. (1986). On the structure of adolescent self-concept. *Journal of Educational Psychology, 78,* 474–481. doi:10.1037/0022-0663.78.6.474

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116–131. doi:10.1037/0022-3514.42.1.116

Candy, P. C. (1991). *Self-direction for lifelong learning: A comprehensive guide to theory and practice*. San-Francisco, CA: Jossey-Bass.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies.* New York, NY: Cambridge University Press. doi:10.1017/CBO9780511571312

Carroll, J. B. (1997). Psychometrics, intelligence and public perceptions. *Intelligence, 24,* 25–52. doi:10.1016/S0160-2896(97)90012-X

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 69–76). New York, NY: Guilford Press.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston, MA: Houghton Mifflin.

*Chamorro-Premuzic, T., Arteche, A., Furnham, A., & Trickot, N. (2009). Assessing pupils' intelligence through self, parental, and teacher estimates. *Educational Psychology, 29,* 83–97. doi:10.1080/01443410802520662

Chamorro-Premuzic, T., & Furnham, A. (2004). A possible model for understanding the personality–intelligence interface. *British Journal of Psychology, 95,* 249–264. doi:10.1348/000712604773952458

*Chamorro-Premuzic, T., & Furnham, A. (2006). Self-assessed intelligence and academic performance. *Educational Psychology, 26,* 769–779. doi:10.1080/01443410500390921

*Chamorro-Premuzic, T., Furnham, A., & Moutafi, J. (2004). The relationship between estimated and psychometric personality and intelligence scores. *Journal of Research in Personality, 38,* 505–513. doi:10.1016/j.jrp.2003.10.002

*Chamorro-Premuzic, T., Moutafi, J., & Furnham, A. (2005). The relationship between personality traits, subjectively-assessed and fluid intelligence. *Personality and Individual Differences, 38,* 1517–1528. doi:10.1016/j.paid.2004.09.018

Chen, M. J., & Chen, H. C. (1988). Concepts of intelligence: A comparison of Chinese undergraduates from Chinese and English schools in Hong Kong. *International Journal of Psychology, 23,* 471–487. doi:10.1080/00207598808247780

*Cogan, L. C., Conklin, A. M., & Hollingworth, H. L. (1915). An experimental study of self-analysis, estimates of associates, and the results of tests. *School & Society, 2,* 171–179.

Cohen, J. (1988). *Statistical power analysis for the social sciences* (2nd ed.). New York, NY: Erlbaum.

Dar-Nimrod, I., & Heine, S. J. (2006, October 20). Exposure to scientific theories affects women's math performance. *Science, 314,* 435. doi:10.1126/science.1131100

*DeNisi, A., & Shaw, J. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology, 62,* 641–644. doi:10.1037/0021-9010.62.5.641

Dobson, K., & Franche, R.-L. (1989). A conceptual and empirical review of the depressive realism hypothesis. *Canadian Journal of Behavioural Science, 21,* 419–433. doi:10.1037/h0079839

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessment of ability. *Journal of Personality and Social Psychology, 57,* 1082–1090. doi:10.1037/0022-3514.57.6.1082

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

Dweck, C. S., & Leggett, E. L. (1988). A social–cognitive approach to motivation and personality. *Psychological Review, 95,* 256–273. doi:10.1037/0033-295X.95.2.256

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. doi:10.1136/bmj.315.7109.629

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84,* 5–17. doi:10.1037/0022-3514.84.1.5

Epstein, S. (1973). The self-concept revisited. *American Psychologist, 28,* 404–416. doi:10.1037/h0034679

Eva, K. W., & Regehr, G. (2007). Knowing when to look it up: A new conception of self-assessment ability. *Academic Medicine, 82*(Suppl. 10), S81–S84. doi:10.1097/ACM.0b013e31813e6755

Eysenck, H. J. (1998). *Intelligence: A new look.* London, England: Transaction.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140. doi:10.1177/001872675400700202

Flugel, J. C. (1947). An inquiry as to popular views on intelligence and related topics. *British Journal of Educational Psychology, 17,* 140–152. doi:10.1111/j.2044-8279.1947.tb02222.x

Freudenthaler, H. H., Spinath, B., & Neubauer, A. C. (2008). Predicting school achievement in boys and girls. *European Journal of Personality, 22,* 231–245. doi:10.1002/per.678

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences, 50,* 723–728. doi:10.1016/j.paid.2010.12.025

Freund, P. A., Kuhn, J.-T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences, 51,* 629–634. doi:10.1016/j.paid.2011.05.033

Froehlich, C. P., & Moser, W. E. (1954). Do counselees remember test scores? *Journal of Counseling Psychology, 1,* 149–152. doi:10.1037/h0061944

Furnham, A. (2000). Parents' estimates of their own and their children's multiple intelligences. *British Journal of Developmental Psychology, 18,* 583–594. doi:10.1348/026151000165869

Furnham, A. (2001). Self-estimates of intelligence: Culture and gender difference in self and other estimates of both general (*g*) and multiple intelligences. *Personality and Individual Differences, 31,* 1381–1405. doi:10.1016/S0191-8869(00)00232-4

Furnham, A. (2005a). Gender and personality differences in self- and other ratings of business intelligence. *British Journal of Management, 16,* 91–103. doi:10.1111/j.1467-8551.2005.00434.x

*Furnham, A. (2005b). Self-estimated intelligence, psychometric intelligence and personality. *Psychologia, 48,* 182–192. doi:10.2117/psysoc.2005.182

*Furnham, A. (2009). The validity of a new, self-report measure of multiple intelligences. *Current Psychology, 28,* 225–239. doi:10.1007/s12144-009-9064-z

*Furnham, A., & Chamorro-Premuzic, T. (2004). Estimating one's own personality and intelligence scores. *British Journal of Psychology, 95,* 149–160. doi:10.1348/000712604773952395

*Furnham, A., & Dissou, G. (2007). The relationship between self-estimated and test-derived scores of personality and intelligence. *Journal of Individual Differences, 28,* 37–44. doi:10.1027/1614-0001.28.1.37

*Furnham, A., & Fong, G. (2000). Self-estimated and psychometrically measured intelligence: A cross-cultural and sex differences study of British and Singaporean students. *North American Journal of Psychology, 2,* 191–200.

Furnham, A., Hosoe, T., & Tang, T. L. (2002). Male hubris and female humility? A cross-cultural study of ratings of self, parental, and sibling multiple intelligence in America, Britain, and Japan. *Intelligence, 30,* 101–115. doi:10.1016/S0160-2896(01)00080-0

*Furnham, A., Kidwai, A., & Thomas, C. (2001). Personality, psychometric intelligence and self-estimated intelligence. *Journal of Social Behavior and Personality, 16,* 97–114.

*Furnham, A., Moutafi, J., & Chamorro-Premuzic, T. (2005). Personality and intelligence: Gender, the Big Five, self-estimated and psychometric intelligence. *International Journal of Selection and Assessment, 13,* 11–24. doi:10.1111/j.0965-075X.2005.00296.x

*Furnham, A., & Rawles, R. (1999). Correlations between self-estimated and psychometrically measured IQ. *Journal of Social Psychology, 139,* 405–410. doi:10.1080/00224549909598400

*Furnham, A., Zhang, J., & Chamorro-Premuzic, T. (2006). The relationship between psychometric and self-estimated intelligence, creativity, personality and academic achievement. *Imagination, Cognition and Personality, 25,* 119–145. doi:10.2190/530V-3M9U-7UQ8-FMBG

Fussell, S. R., & Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology, 21,* 445–454. doi:10.1002/ejsp.2420210507

*Gabriel, M. T., Critelli, J. W., & Ee, J. S. (1994). Narcissistic illusions in self-evaluations of intelligence and attractiveness. *Journal of Personality, 62,* 143–155. doi:10.1111/j.1467-6494.1994.tb00798.x

Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. *Adult Education Quarterly, 48,* 18–33. doi:10.1177/074171369704800103

Gati, I., Noa, S., & Krausz, M. (2001). "Should I use a computer-assisted career guidance system?" It depends on where your career decision-making difficulties lie. *British Journal of Guidance & Counseling, 29,* 301–321.

Goethals, G. R., & Klein, W. M. P. (2000). Interpreting and inventing social reality: Attributional and constructive elements in social comparison. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 23–44). New York, NY: Kluwer Academic/Plenum Press. doi:10.1007/978-1-4615-4237-7_2

Goff, M., & Ackerman, P. L. (1992). Personality–intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology, 84,* 537–552. doi:10.1037/0022-0663.84.4.537

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6,* 48–60. doi:10.1177/1745691610393521

Gottfredson, L. S. (1997a). Mainstream science on intelligence. *Intelligence, 24,* 13–23. doi:10.1016/S0160-2896(97)90011-8

Gottfredson, L. S. (1997b). Why *g* matters: The complexity of everyday life. *Intelligence, 24,* 79–132. doi:10.1016/S0160-2896(97)90014-3

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52,* 1115–1124. doi:10.1037/0003-066X.52.10.1115

Greven, C. U., Harlaar, N., Kovas, Y., Chamorro-Premuzic, T., & Plomin, R. (2009). More than just IQ: School achievement is predicted by self-perceived abilities—but for genetic rather than environmental reasons. *Psychological Science, 20,* 753–762. doi:10.1111/j.1467-9280.2009.02366.x

Guenther, C. L., & Alicke, M. D. (2010). Deconstructing the better-than-average effect. *Journal of Personality and Social Psychology, 99,* 755–770. doi:10.1037/a0020959

Guilford, J. P. (1982). Cognitive psychology's ambiguities: Some suggested remedies. *Psychological Review, 89,* 48–59. doi:10.1037/0033-295X.89.1.48

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement, 48,* 1–4. doi:10.1177/001316448804800102

Guimond, S. (Ed.). (2006). *Social comparison and social psychology.* Cambridge, England: Cambridge University Press.

Guimond, S., Branscombe, N. R., Brunot, S., Buunk, A. P., Chatard, A., Désert, M., . . . Yzerbyt, V. (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology, 92,* 1118–1134. doi:10.1037/0022-3514.92.6.1118

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92,* 373–385. doi:10.1037/0021-9010.92.2.373

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* Cambridge, England: Cambridge University Press.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106,* 766–794. doi:10.1037/0033-295X.106.4.766

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–83. doi:10.1017/S0140525X0999152X

Hirschi, A., & Läge, D. (2008). Using accuracy of self-estimated interest type as a sign of career choice readiness in career assessment of secondary students. *Journal of Career Assessment, 16,* 310–325. doi:10.1177/1069072708317372

*Hodgson, M. L., & Cramer, S. H. (1977). The relationship between selected self-estimated and measured abilities in adolescents. *Measurement & Evaluation in Guidance, 10,* 98–105.

Hogan, H. (1978). IQ self-estimates of males and females. *Journal of Social Psychology, 106,* 137–138. doi:10.1080/00224545.1978.9924160

*Holling, H., & Preckel, F. (2005). Self-estimates of intelligence—methodological approaches and gender differences. *Personality and Individual Differences, 38,* 503–517. doi:10.1016/j.paid.2004.05.003

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Intelligence and its measurement: A symposium. (1921). *Journal of Educational Psychology, 12,* 123–147. doi:10.1037/h0076078

Jackson, L. A., Hodge, C. N., & Ingram, J. M. (1994). Gender and self-concept: A reexamination of stereotyping differences and the role of gender attitudes. *Sex Roles, 30,* 615–630. doi:10.1007/BF01544666

Jussim, L., Yen, H., & Aiello, J. R. (1995). Self-consistency, self-enhancement, and accuracy in reactions to feedback. *Journal of Experimental Social Psychology, 31,* 322–356. doi:10.1006/jesp.1995.1015

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate linear model for meta-analysis. *Psychological Methods, 1,* 227–235. doi:10.1037/1082-989X.1.3.227

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender–math stereotypes influence women's susceptibility to stereotype threat. *Journal of Experimental Social Psychology, 43,* 825–832. doi:10.1016/j.jesp.2006.08.004

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

*Kornilova, T. V., Kornilov, S. A., & Chumakova, M. A. (2009). Subjective evaluations of intelligence and academic self-concept predict academic achievement: Evidence from a selective student population. *Learning and Individual Differences, 19,* 596–608. doi:10.1016/j.lindif.2009.08.001

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82,* 180–188. doi:10.1037/0022-3514.82.2.180

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77,* 221–232. doi:10.1037/0022-3514.77.2.221

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134. doi:10.1037/0022-3514.77.6.1121

Kruger, J., & Dunning, D. (2002). Unskilled and unaware—but why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology, 82,* 189–192. doi:10.1037/0022-3514.82.2.189

Kruglanski, A. W., & Mayseless, O. (1990). Classic and current social comparison research: Expanding the perspective. *Psychological Bulletin, 108,* 195–208. doi:10.1037/0033-2909.108.2.195

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95,* 179–188. doi:10.1037/0033-2909.95.2.179

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435–447.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86,* 148–161. doi:10.1037/0022-3514.86.1.148

Lecky, P. (1945). *Self-consistency: A theory of personality.* New York, NY: Island Press.

Lesko, A. C., & Corpus, J. H. (2006). Discounting the difficult: How high math-identified women respond to stereotype threat. *Sex Roles, 54,* 113–125. doi:10.1007/s11199-005-8873-2

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183. doi:10.1016/0030-5073(77)90001-0

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149–171. doi:10.1016/0030-5073(80)90052-5

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92,* 1672–1682. doi:10.1037/0021-9010.92.6.1672

Lohman, D. F. (2001). Issues in the definition and measurement of abilities. In J. M. Collis & S. J. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 79–98). Mahwah, NJ: Erlbaum.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology, 51,* 405–444. doi:10.1146/annurev.psych.51.1.405

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280–296. doi:10.1037/0021-9010.67.3.280

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2,* 77–172. doi:10.1007/BF01322177

Martin, R. (2000). "Can I do X?": Using the proxy model to predict performance. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison* (pp. 67–80). New York, NY: Plenum Press. doi:10.1007/978-1-4615-4237-7_4

Maxwell, N., & Lopus, J. (1994). The Lake Wobegon effect in student self-report data. *The American Economic Review, 84,* 201–205.

Mayseless, O., & Kruglanski, A. W. (1987). Accuracy of estimates in the social comparison of abilities. *Journal of Experimental Social Psychology, 23,* 217–229. doi:10.1016/0022-1031(87)90033-3

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136–181). New York, NY: Guilford Press.

McMahon, M., & Patton, W. (2002). Using qualitative assessment in career counselling. *International Journal for Educational and Vocational Guidance, 2,* 51–66. doi:10.1023/A:1014283407496

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525–543. doi:10.1007/BF02294825

Meyer, W.-U. (1982). Indirect communications about perceived ability estimates. *Journal of Educational Psychology, 74,* 888–897. doi:10.1037/0022-0663.74.6.888

Miflin, B. M., Campbell, C. B., & Price, D. A. (2000). A conceptual framework to guide the development of self-directed, lifelong learning in problem-based medical curricula. *Medical Education, 34,* 299–306. doi:10.1046/j.1365-2923.2000.00564.x

Morse, S., & Gergen, K. J. (1970). Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology, 16,* 148–156. doi:10.1037/h0029862

Mussweiler, T. (2003a). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110,* 472–489. doi:10.1037/0033-295X.110.3.472

Mussweiler, T. (2003b). "Everything is relative": Comparison processes in social judgment. *European Journal of Social Psychology, 33,* 719–733. doi:10.1002/ejsp.169

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101. doi:10.1037/0003-066X.51.2.77

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88,* 622–637. doi:10.1037/0033-2909.88.3.622

Nevo, B., & Khader, A. (1995). Cross-cultural, gender, and age differences in Singaporean mothers' conceptions of children's intelligence. *Journal of Social Psychology, 135,* 509–517. doi:10.1080/00224545.1995.9712219

Ng, J., & Earl, J. (2008). Accuracy in self-assessment: The role of ability, feedback, self-efficacy and goal orientation. *Australian Journal of Career Development, 17,* 39–50.

Nisbett, R. (2003). *The geography of thought.* New York, NY: Free Press.

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

Oswald, D. L., & Harvey, R. D. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology, 19,* 338–356. doi:10.1007/s12144-000-1025-5

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

*Paulhus, D. L., Lysy, D. C., & Yik, M. S. M. (1998). Self-reported measures of intelligence: Are they still useful as proxy IQ tests? *Journal of Personality, 66,* 525–554. doi:10.1111/1467-6494.00023

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90,* 175–181. doi:10.1037/0021-9010.90.1.175

Prediger, D. J. (1999). Basic structure of work-relevant abilities. *Journal of Counseling Psychology, 46,* 173–184. doi:10.1037/0022-0167.46.2.173

*Proyer, R. T., & Ruch, W. (2009). Intelligence and gelotophobia: The relations of self-estimated and psychometrically measured intelligence to the fear of being laughed at. *Humor, 22,* 165–181. doi:10.1515/HUMR.2009.008

Rammstedt, B., & Rammsayer, T. H. (2000). Sex differences in self-estimates of different aspects of intelligence. *Personality and Individual Differences, 29,* 869–880. doi:10.1016/S0191-8869(99)00238-X

*Rammstedt, B., & Rammsayer, T. H. (2002a). Die Erfassung von selbsteingeschätzter Intelligenz: Konstruktion, teststatistische Überprüfung und erste Ergebnisse des ISI [Assessment of self-estimated intelligence: Construction, statistical testing, and first results of the Inventory of Self-Estimated Intelligence (ISI)]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23,* 435–446. doi:10.1024//0170-1789.23.4.435

*Rammstedt, B., & Rammsayer, T. H. (2002b). Self-estimated intelligence: Gender differences, relationship to psychometric intelligence and moderating effects of level of education. *European Psychologist, 7,* 275–284. doi:10.1027//1016-9040.7.4.275

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2009). HLM 6.08: Hierarchical linear and nonlinear modeling [Computer software]. Chicago, IL: Scientific Software International.

*Reilly, J., & Mulhern, G. (1995). Gender differences in self-estimated IQ: The need for care in interpreting group data. *Personality and Individual Differences, 18,* 189–192. doi:10.1016/0191-8869(94)00148-L

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566. doi:10.1037/0033-2909.114.3.552

Rosenberg, F. R., & Simmons, R. G. (1975). Sex differences in the self-concept in adolescence. *Sex Roles, 1,* 147–159. doi:10.1007/BF00288008

Ruzgis, P., & Grigorenko, E. L. (1994). Cultural meaning systems, intelligence and personality. In R. J. Sternberg & P. Ruzgis (Eds.), *Personality and intelligence* (pp. 248–270). Cambridge, England: Cambridge University Press.

Sampson, J. (1994). Factors influencing the effective use of computer-assisted career guidance: The North American experience. *British Journal of Guidance & Counseling, 22,* 91–106.

Sampson, J., & Watts, A. G. (1992). Computer-assisted career guidance systems and organizational change. *British Journal of Guidance & Counseling, 20,* 328–343.

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. doi:10.1037/0033-2909.124.2.262

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the work of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162–173. doi:10.1037/0022-3514.86.1.162

Segall, M. H., Lonner, W. J., & Berry, J. W. (1998). Cross-cultural psychology as a scholarly discipline: On the flowering of culture in behavioral research. *American Psychologist, 53,* 1101–1110. doi:10.1037/0003-066X.53.10.1101

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46,* 407–441.

Shokar, G. S., Shokar, N. K., Romero, C. M., & Bulik, R. J. (2002). Self-directed learning: Looking at outcomes with medical students. *Medical Student Education, 34,* 197–200.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's *z* transformation be used? *Journal of Applied Psychology, 72,* 146–148. doi:10.1037/0021-9010.72.1.146

Skaalvik, S., & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50,* 241–252. doi:10.1023/B:SERS.0000015555.40976.e6

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research, 1,* 293–319. doi:10.1023/A:1009932704458

Spearman, C. E. (1927). *The abilities of man: Their nature and measurement.* New York, NY: Macmillan.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629. doi:10.1037/0003-066X.52.6.613

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811. doi:10.1037/0022-3514.69.5.797

*Steinmayr, R., & Spinath, B. (2009). What explains boys' stronger confidence in their intelligence? *Sex Roles, 61,* 736–749. doi:10.1007/s11199-009-9675-8

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge, England: Cambridge University Press.

Sternberg, R. J. (2000). The concept of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 3–15). Cambridge, England: Cambridge University Press.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41,* 37–55. doi:10.1037/0022-3514.41.1.37

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence? Contemporary viewpoints on its nature and definition.* Norwood, NJ: Ablex.

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49,* 479–502. doi:10.1146/annurev.psych.49.1.479

Stone, E. R., Dodrill, C. L., & Johnson, N. (2001). Depressive cognition: A test of depressive realism versus negativity using general knowledge questions. *Journal of Psychology: Interdisciplinary and Applied, 135,* 583–602. doi:10.1080/00223980109603722

Swann, W. B., Griffin, J., Predmore, S., & Gaines, B. (1987). The cognitive–affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology, 52,* 881–889. doi:10.1037/0022-3514.52.5.881

Szymanowicz, A., & Furnham, A. (2011). Gender differences in self-estimates of general, mathematical, spatial and verbal intelligence: Four meta analyses. *Learning and Individual Differences, 21,* 493–504. doi:10.1016/j.lindif.2011.07.001

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210. doi:10.1037/0033-2909.103.2.193

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago, IL: University of Chicago Press.

Trope, Y. (1986). Self-enhancement and self-assessment in achievement behavior. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (Vol. 2, pp. 350–378). New York, NY: Guilford Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. doi:10.1177/109442810031002

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31,* 33–51. doi:10.1177/0022022100031001004

*Visser, B. A., Ashton, M. C., & Vernon, P. A. (2008). What makes you think you're so smart? Measured abilities, personality, and sex differences in relation to self-estimates of multiple intelligences. *Journal of Individual Differences, 29,* 35–44. doi:10.1027/1614-0001.29.1.35

von Stumm, S., Chamorro-Premuzic, T., & Furnham, A. (2009). Decomposing self-estimates of intelligence: Structure and sex differences across 12 nations. *British Journal of Psychology, 100,* 429–442. doi:10.1348/000712608X357876

*Webb, W. B. (1955). Self-evaluations, group evaluations, and objective measures. *Journal of Consulting Psychology, 19,* 210–212. doi:10.1037/h0047190

*Wells, L. E., & Sweeney, P. D. (1986). A test of three models of bias in self-assessment. *Social Psychology Quarterly, 49,* 1–10. doi:10.2307/2786852

*Westbrook, B. W., Buck, R. W., Wynne, D. C., & Sanford, E. (1994). Career maturity in adolescence: Reliability and validity of self-ratings of abilities by gender and ethnicity. *Journal of Career Assessment, 2,* 125–161. doi:10.1177/106907279400200203

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696–716. doi:10.1037/0022-3514.89.5.696

Wilgenbusch, T., & Merrell, K. W. (1999). Gender differences in self-concept among children and adolescents: A meta-analysis of multidimensional studies. *School Psychology Quarterly, 14,* 101–120. doi:10.1037/h0089000

*Wolff, R., & Wasden, R. (1969). Measured intelligence and estimates by nursing instructors and nursing students. *Psychological Reports, 25,* 77–78. doi:10.2466/pr0.1969.25.1.77

Wolman, S. D. (2009). *Self-estimates of job performance and learning potential* (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.

Young, R. A., & Collin, A. (2004). Introduction: Constructivism and social constructionism in the career field. *Journal of Vocational Behavior, 64,* 373–388. doi:10.1016/j.jvb.2003.12.005

Received November 8, 2010
Revision received September 30, 2011
Accepted November 3, 2011 ■