Review of Educational Research
Spring 2005, Vol. 75, No. 1, pp. 63–82

The Validity of Self-Reported Grade Point Averages, Class Ranks, and Test Scores: A Meta-Analysis and Review of the Literature

Nathan R. Kuncel, Marcus Credé, and Lisa L. Thomas
University of Illinois, Urbana-Champaign

Self-reported grades are heavily used in research and applied settings because of the importance of grades and the convenience of obtaining self-reports. This study reviews and meta-analytically summarizes the literature on the accuracy of self-reported grades, class ranks, and test scores. Results based on a pairwise sample of 60,926 subjects indicate that self-reported grades are less construct valid than many scholars believe. Furthermore, self-reported grade validity was strongly moderated by actual levels of school performance and cognitive ability. These findings suggest that self-reported grades should be used with caution. Situations in which self-reported grades can be employed more safely are identified, and suggestions for their use in research are discussed.

Keywords: GPA, grade point average, rank, validity.

The grade point average (GPA) is one of the most studied variables in education and educational psychology. Unfortunately, it is often impractical or even impossible to obtain school transcripts. Instead, self-reported grade point averages are often obtained for the purposes of research or to facilitate placement of students. An ongoing concern has been the extent to which self-reported grades are an accurate reflection of actual earned grades. The purpose of this study was to review the literature on self-reported grades and grade point averages and to conduct meta-analyses of the correlation and average differences between self-reported grades and actual earned grades, as well as the base rates of under-reporting, accurate reporting, and over-reporting of grade point averages. This study also aggregates and clarifies the literature on the moderating effect of other variables on the construct validity of self-reported grades (e.g., current class rank, gender, race, general cognitive ability). This synthesis will provide data of fundamental importance to researchers and practitioners who use self-reported grades in their research and work.

The common use of self-reported grades is understandable, because grade point averages are important. Not only are they summaries of student learning, they are also important predictors of performance at other levels of education and of other important life outcomes. For example, high school GPA is one of the best predictors of college grades (Ramist, 1984; Willingham & Breland, 1982), and college GPA is a robust predictor of performance in graduate school (Kuncel, Hezlett, & Ones, 2001), pharmacy programs (Kuncel, Credé, Thomas, Klieger, Seiler, & Woo, in press), business school (Kuncel, Credé, & Thomas, 2004), and law school (Linn & Hastings, 1984).

Grades are also heavily used for making hiring decisions about students who are just out of college. Albrecht, Carpenter, and Sivo (1994) surveyed 664 professional recruiters, and over 80% of them reported using a minimum GPA cutoff in their decision-making. The emphasis on grades in hiring decisions is not without merit, as college grade point averages are also valid predictors of job performance (Roth, Be Vier, Switzer, & Schippman, 1996) and salary (Roth & Clarke, 1998).

Given the importance of GPA, it is also important to know if self-reported GPA can be substituted for actual GPA for the purposes of, at the very least, research in these areas as well as for decision-making (e.g., academic counseling, hiring decisions). The validity of self-reported grades has implications for educational researchers, student placement, and personnel selection decisions. Some researchers have held generally negative views of using self-reported GPA (e.g., Goldman, Flake, & Matheson, 1990), while others have argued in favor of using self-reported GPA (e.g., Cassady, 2001), arguing that the relationship, although imperfect, is close enough for research and practical purposes. Two related questions need to be answered to address this problem. The first is the magnitude of the discrepancy between self-reported and actual grades. The second is the nature or cause of the discrepancy.

Error in Self-Reported Grades

The magnitude and cause of the discrepancy speak to the construct validity of self-reported grades, which is the fundamental concern of this study. In considering the construct validity of self-reported grades, two general and related questions can be posed. The first is, To what extent are self-reported grades accurate measures of actual earned grades? The second is, To what extent do self-reported grades reflect the learning, ability, persistence, achievement, and whatever else we believe (or perhaps hope) that actual grades reflect? Because the most common use of self-reported grades is as a direct substitute for actual grades, the primary objective of this study is to marshal all of the available evidence to provide the best possible answer to the first question, while also providing additional information to inform the second question. That is, we accumulate evidence to determine if self-reported grades can be substituted for actual grades. To evaluate this evidence, we need to consider that an imperfect relationship between self-reported and actual grades can be due to either random error variance or systematic variance that is unrelated to actual school-reported grades.

An imperfect correlation can indicate two things. The less serious situation would be that self-reported GPAs simply have had some amount of random error added to them. For example, an individual's overall GPA is likely to be subject to imperfect storage and retrieval from memory. If this results in random error, then the self-reported GPA can be used with the small price of being a somewhat less reliable measure. For research purposes, this will result in attenuated effects and lower power (Humphreys & Drasgow, 1989). For decision-making in educational settings, this will increase the number of incorrect decisions (e.g., more students will be placed in courses that are not optimal for their current skill level). The presence of random error in self-reported grades does not necessarily make their use inappropriate, as the purpose for which the measure will be used is critical for making judgments about construct validity evidence (Messick, 1989). For research purposes, the presence of random error is less of a concern, so long as researchers are aware of it

and take appropriate steps (e.g., conducting power analyses during study design, correcting for attenuation). For applied purposes, the utility of less accurate but more convenient self-reported grades would need to be weighed against the increased rate of inaccurate decisions. Assuming that the amount of additional random error variance is not large, the work of both scientists and practitioners will not be seriously affected.

The more serious situation for both research and practice would be the presence of systematic sources of invalid variance. In this case, the imperfect correlation is the result of other sources of variance beyond (or in addition to) random error. Such systematic variance has been termed *construct-irrelevant variance* (Messick, 1995). It is critical to note that this term refers to reliable variance that is not related to the construct of interest. In Messick's words, a measure with construct irrelevant variance is found to be "containing excess reliable variance associated with other distinct constructs" (Messick, 1995, p. 742). For research purposes, it cannot be assumed that results obtained with self-reported grades are simply attenuated. Depending on the nature of the other sources of variance, effects could be increased or decreased, potentially taking on any value in the range of possible values for a given correlation between two variables and their relationship with a third (McCornack, 1956). In other words, the correlation that would be obtained with actual GPA could be higher or lower than what is observed with self-reported GPA. This could, of course, lead to misleading research conclusions.

Ironically, despite any observed inaccuracy in the relationship between self-reported grades and actual grades, it is possible that self-reported grades possess superior construct validity as a measure of other constructs, including learning, motivation, ability, and achievement. Although this seems unlikely to us, given the data that we have analyzed, it is still possible that student self-reports better reflect those other constructs. Students might, for example, mentally adjust their prior grades to reflect their true capabilities. Again, this seems unlikely, but it helps to illustrate how we can consider two questions for the validity of self-reported grades: the extent to which they measure actual grades, and the extent to which they measure the other constructs for which researchers and practitioners use grades as operationalizations.

For applied purposes, the presence of systematic invalid variance could decrease or even increase the number of correct decisions. For example, the self-reported GPA might be a better reflection of student skill levels. Alternatively, self-reported GPA may partially reflect intentional deception that is negatively related to subsequent success in school. Again the second scenario seems more likely, particularly in light of our meta-analytic results, yet the applied implications of using self-reported GPA fundamentally depend on the nature of the other sources of variance.

Evidence for the presence of systematic invalid variance can come from several different sources. Two related pieces of evidence are a consistent unequal pattern of over-reporting or under-reporting of GPAs and a consistent mean difference between self-reported and actual school-reported grades. If, on average, more students over-report their grades, then the difference is not random. Similarly, a consistent mean difference between self-reported and school-reported grades also suggests non-random errors. Of course, if the difference is a constant across all individuals (e.g., everyone inflates their GPA by half a point), then the self-reported GPA would be acceptable for correlational research and decision-making, except at the extreme ends where scores could be truncated by the limit of the scale (e.g., one cannot have a 4.3 GPA on a 0.0–4.0 scale). However, it is unlikely that inaccurate reporting is a constant across people, and it is far more likely that individual difference (e.g., self-monitoring, socialization) and situation factors (e.g., benefits of misleading or costs of being caught) may yield different amounts of inaccuracy. This third variable would be a source of construct irrelevant variance. The existence of a third variable that moderates the relationship between self-reported and school-reported GPA would be clear evidence that the differences between self-reported GPA and actual GPA are not purely random error.

Reliability of Self-Reported Grades

For readers who are familiar with classical test theory, we can also consider the correlation between self-reported grades and actual grades from records to be an estimate of the reliability of self-reported grades. If we define reliability (ρ_{xx}) to be the population correlation between observed scores and true scores, or $\rho_{xx} = \rho_{(\text{observed scores, true scores})}$, and if we view grades from records to be true scores, then the observed correlation between observed self-reported grades and school-reported grades is an estimate of reliability.

Traditionally, reliability estimates are based on the correlation between two observed scores for the same, or effectively the same (i.e., parallel), measures. These estimates include test–retest, alternate forms, and coefficients of internal consistency. In the current study, it is not necessary to base estimates on two sets of observed scores (e.g., the correlation between self-reported grades a week apart), because a true score measure actually exists. Simultaneously considering the effects of moderators on these reliability estimates can be viewed as examining reliability within the framework of Generalizability Theory (see Crocker and Algina, 1986 for a review).

Self-Reported Grade Validity: An Illustration

In Figure 1, we illustrate a hypothetical but plausible portioning of variance in self-reported grades. The bulk of the variance in self-reported grades is assumed to be valid variance associated with actual grades. However, the second largest portion of variance is assumed to be systematic invalid variance. This is variability in self-reported grades that is caused by other factors or variables but is not associated with self-reported grades. It is this portion of the total variance that is most troubling because it may cause spurious relationships with other variables. Finally, a modest portion of the total variance in self-reported grades is assumed to be random error. By definition, this portion of the variance is not associated with anything (being random). Any correlations obtained between self-reported grades and other variables will be attenuated because self-reported grades are simply noisier measures of grade point averages than actual school-reported grade point averages. Previous research provides pieces of information for each of these issues though a variety of analyses.

Previous Research

Previous research on self-reported grade point averages has generally examined three different aspects of the relationship between self-reported GPA and actual earned GPA. The most common connection is a correlation between self-reported and actual GPAs. Consistent rank order differences between self-reported and actual grades creates a situation where research results obtained from self-reported grades may not be the same as what would be obtained with actual GPA. The correlation between the two has ranged from near unity to unacceptably low values. Correlations reported in the literature for overall GPA or GPA in broad discipline areas range from a low of .45 (Jung & Moore, 1970) to a high of .98 (Kirk & Sereda, 1969). The extent to which differences are due to sampling error or the influence of individual differences or situational factors is not currently known. Resolving that question is one of the objectives of this study.

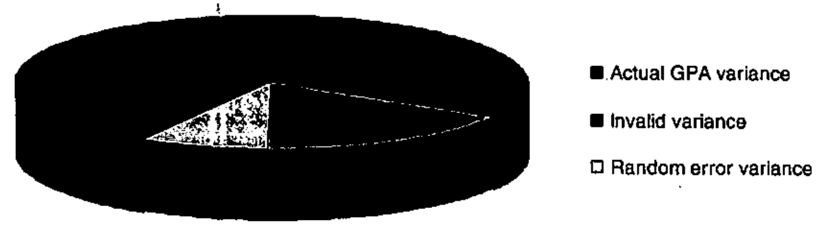

FIGURE 1. *Hypothetical breakdown of variance in self-reported GPA. Actual GPA variance = 75%, invalid variance = 17%, random error variance = 8%.*

The second most common information is the percentage of students who reported accurately, over-reported their GPA, and under-reported their GPA. These data are particularly useful for a general understanding of the extent to which inaccurate reports can be treated as random error or as due to some systematic bias. Related to over-reporting and under-reporting base rates are estimates of the average mean difference between self-reported and school-reported grades. If responding errors were random, we would expect the average mean difference to be zero. Across studies it appears that, on average, students over-report their GPAs. The percentage of student's over-reporting is often over twice as large as the percentage of under-reporting (Bahrick, Hall, & Berger, 1996), sometimes reaching ratios as large as 12 to 1 (Freeberg, 1988) or even 48 to 1 (Zimmerman, Caldwell, & Bernat, 2002). Although this phenomenon appears to be nearly ubiquitous, what remains unclear is whether gender or other demographic or individual difference variables are related to different amounts of over-reporting or under-reporting.

Not surprisingly, the average mean difference between self-reported and actual GPA can be large, *d* values of over .50 being common (e.g., Richards & Lutz, 1967a). Also, the evidence clearly indicates that inaccuracy is not the same across all students. That is, only a subset of all students misrepresent their grades. For example, Maxey and Ormsby (1971) found that 2% of students report a grade that deviated from the school-reported grade by an entire letter grade, while 22% were inaccurate to a lesser extent.

There is also evidence that that those students who tend to be inaccurate differ from their more accurate peers on several dimensions. That is, inaccuracy is predictable to some extent, as certain individual difference variables appear to moderate the reliability of self-reported grades.

Moderators of Self-Reported GPA Reliability

Investigations of moderators of self-reported GPA correlations with actual GPA have focused on the effects of actual GPA, ability, and personality. In general, this research has proposed that those with lower grades, lower ability, or certain personality traits will be more likely to intentionally misrepresent themselves when reporting their grades or test scores. Several scholars have consistently found that actual GPA moderates the reliability of self-reported GPA, such that students with lower actual GPA tend to report their grades less reliably. The methods used to examine this question have varied from study to study. Some have calculated correlations with GPA bands, whereas others examine percentages of over- and under-reporting at different grade cutoffs. For example, Dunnette (1952) found that students with a school-reported average below C were more likely to misrepresent their grades than students with a higher average. Dobbins, Farh, and Werbel (1993) also reported that the correlation between self-reported grades and actual school-reported grades was higher for students who actually had higher grade point averages.

Finally, some scholars have examined the moderating effect of individual difference variables on the accuracy of self-reported grades. Maxey and Ormsby (1971) reported that the correlation between actual and self-reported grades increased with student ACT scores. Similar findings were reported by Schiel and Noble (1991) and ACT (1973). These results suggest that grades are more accurately reported by students with higher ability levels. There is less information about personality moderators of self-reported grades, but self-monitoring also appears to have a moderating effect on inflated reporting of grades (Dobbins, Farh, & Werbel, 1993). Self-monitors tend to have a stronger attention to and willingness to engage in impression management. This construct, developed by Snyder (1987, 1974), holds promise for helping to explain inaccuracies in self-reported grades. Very few studies have examined moderators other than demographic, ability, and actual achievement variables. Additional research in this area would greatly inform the use of self-reported grades and, more broadly, research on self-reports.

Overall, researchers who report investigating the existence of moderating effects have found evidence to support them. To the extent this is true, it can have substantial implications for the use of self-reported grades for both applied (e.g., counseling) and research purposes, because the self-reported grade is not purely a less accurate measure of grades. Rather, it is also affected by other variables. Determining the overall magnitude of the reliability of grades, establishing the presence or absence of moderating effects, and identifying situations where self-reported grades can be used with greater confidence are the three overarching goals of this study.

Methods

The Hunter and Schmidt (1990, 2004) psychometric meta-analytic method was used in this study to aggregate effects across primary studies. This method has several desirable features, including the ability to account for variability across effect sizes due to sampling error and other statistical artifacts. Data were coded into a database from primary studies and then sorted into like analyses (e.g., correlations between self-reported GPAs and actual GPAs for high school students). A program developed by Schmidt, Hunter, Viswesvaran, and colleagues was used for data analysis for each meta-analysis reported in this study. This program applies the interactive meta-analytic procedure with other refinements that improve the precision of the original method (Law, Schmidt, & Hunter, 1994a, 1994b; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1990).

The Hunter and Schmidt method was originally developed within industrial/organizational psychology to conduct meta-analyses of validity studies that used correlational analyses and was therefore a particularly appropriate method for the current study. For the two top journals in industrial/organizational psychology, the Hunter and Schmidt method has been used in 91 published peer-reviewed studies. Examinations of the accuracy of the method have also been conducted. Burke, Raju, and Pearlman (1986) examined five meta-analytic procedures, including the Hunter and Schmidt method. They conducted empirical comparisons across the methods, found that they all yielded similar estimates, and concluded that "the five validity generalization procedures will lead to the same general conclusions" (p. 349). More recently, Schulze (2004) conducted Monte-Carlo simulation comparisons across a number of methods and concluded that the Hunter and Schmidt method yielded accurate results.

There are three major goals in a meta-analysis. The first is to produce an estimate of the average effect size. The second is to estimate the extent to which the variability observed in the literature is due to real effects instead of statistical artifacts. The most commonly considered statistical artifacts are sampling error, differential measurement reliability, differential dichotomization, and, where appropriate, range restriction. These artifacts can make studies appear to have different results when, in fact, the true effect does not vary. The Hunter and Schmidt method allows researchers to take these effects into account. To estimate the average effect, a sample-size-weighted average observed correlation is calculated. This is the best point estimate for the average effect across studies. The Hunter and Schmidt method produces some additional useful statistics. First, k is the number of studies and N is the total sample across the studies combined in that analysis. Next, is the SD_{obs}, which is the observed standard deviation of the correlations. This statistic quantifies the extent to which effects vary in the literature without accounting for any artifacts. The estimate of SD_{rho} is the standard deviation of true score correlations. This is an estimate of the amount of variability across studies that remains after accounting for artifacts.

In the current study, it was appropriate to account for dichotomization, reliability, and sampling error. No studies utilized dichotomization, and the reliability of school-reported grades was assumed to be effectively the same and at unity across studies. Therefore, only sampling error was addressed. If SD_{rho} is large, it is reasonable to conclude that one or more substantive moderators may operate that are producing the variability. If SD_{rho} is nearly zero, then the variability observed in the literature is largely artifactual and there is good evidence that strong moderators are not present. Finally, we present the 90% credibility interval. The 90% credibility interval is not a confidence interval. It provides a range of plausible values that could result from unexamined moderators. The credibility interval is based on

SD_{rho} and not on the standard error of the correlation. In contrast, a confidence interval is based on the standard error, which is largely a function of sample size. Therefore, confidence intervals reflect sampling error, while the credibility interval reflects variance in effects across studies that cannot be accounted for with artifacts including sampling error. Therefore, this remaining variability is attributed to unaddressed and possibly unknown moderator variables. Because meta-analyses typically have very large sample sizes, the confidence intervals are generally very small. In contrast, the credibility interval is not dependent on sample size and can be quite large. When credibility intervals are large, we can conclude that the effect in question may truly vary across situations or samples. This is a good indication that examining moderators is appropriate.

Description of the Meta-Analytic Database

Numerous sources were searched for studies examining the relationship between self-reported and actual grade point averages and class ranks. Electronic searches were performed on PsycINFO (1872–2003), ERIC (1966–2003), and Dissertation Abstracts (1861–2003). We also searched listings of technical reports that are available from American College Testing (ACT), The College Board, and the Educational Testing Service (ETS). Within every article, report, or dissertation we obtained, we examined the reference list for other promising articles or reports that were not identified in our original electronic searches. Finally, we did Social Science Citation Index searches for a few key articles that were commonly cited to identify other studies that examined the same research question.

Each article was coded by one of the three authors. Information captured from each article included the nature of the GPA or rank (e.g., college, high school), the effect sizes, sample sizes, and moderator information including study conditions, race, gender, and time lag between actual and self-reported information. Precise coding and combination of data are critical for the production of a meta-analysis. If data examining fundamentally different samples or variables are unintentionally combined, it may jeopardize the findings. The result would be a mixing of potentially different studies that could yield an uninterpretable blend. Stated simply, if the goal is to conduct a meta-analysis on oranges, we need to make sure that only oranges and no apples are included. If the goal is to study fruit, we are fine with the apples and the oranges but need to keep the vegetables out. The composition of studies included in any given analysis should depend on the research question.

To help ensure the reliability and validity of this study, a series of procedures was used. First, a standardized coding sheet was developed to ensure systematic collection of all relevant information. Second, to produce as comprehensive a study as possible, effect sizes were computed from available information (e.g., raw data, frequency tables) when studies did not report data in a standard effect size (i.e., d or r) format. We also sought out high quality unpublished data to improve the comprehensiveness of the study. Third, it is important to avoid including research with overlapping or identical samples. Data were sorted by sample size, effect size, and author name to help address this problem. In some cases, it appears that multiple authors reported identical data. In all cases with overlapping data, the largest and most complete study was retained and the smaller overlapping studies were excluded.

Several studies have found a high reliability for the information coded in meta-analysis (Kuncel, Hezlett, & Ones, 2001; Whetzel & McDaniel, 1988; Zakzanis,

1998). However, two additional precautions were taken to ensure coding accuracy in this study. First, all coded data were examined by one other coder and proofed for accuracy. It was particularly important to make sure that two coders agreed on the classification of study findings into different potential primary or moderator analyses. Second, the first author, who has the most experience conducting meta-analyses, proofed a small random sample of the overall set of articles.

Data were graphed for continuous moderators (i.e., cognitive ability and actual academic performance) by using the information from those studies that presented reliabilities at different bands of actual academic performance (e.g., reliability for grades of D or lower versus C or better) and cognitive ability levels. These data are presented in Figure 2 and Figure 3, respectively.

The final database included 37 independent samples with 158 effect sizes across 60,926 subjects. The total effect size estimate includes both standardized mean differences and correlations. No analysis included multiple effect sizes from the same sample, and independence was not violated.

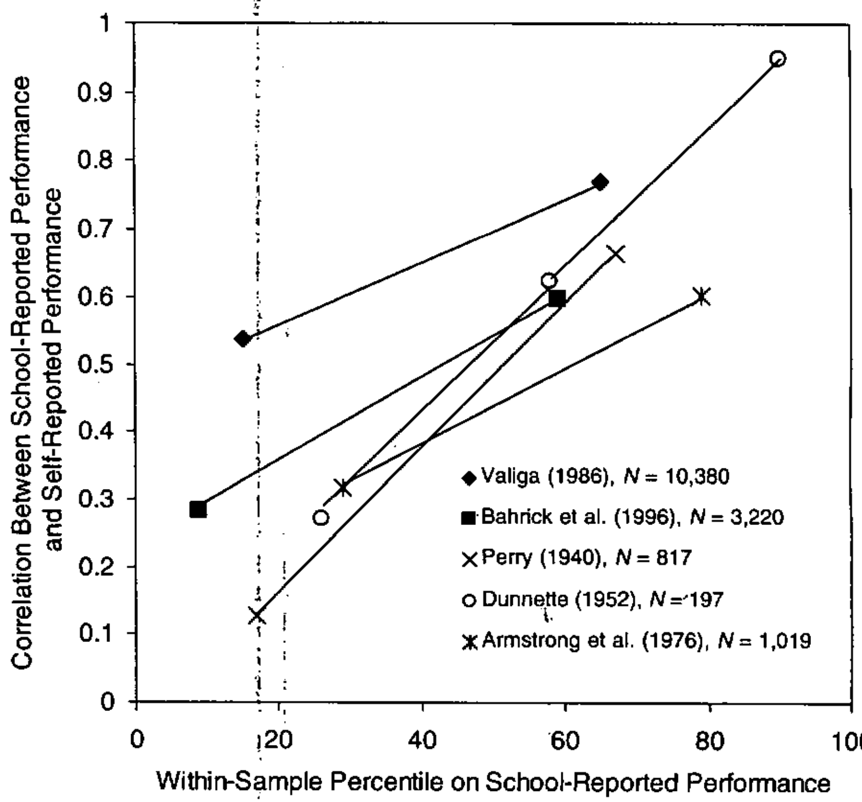

FIGURE 2. *Self-reported GPA reliability at various levels of school-reported performance. The sample size for Bahrick et al. (1996) is the number of grades reported rather than the number of students.*

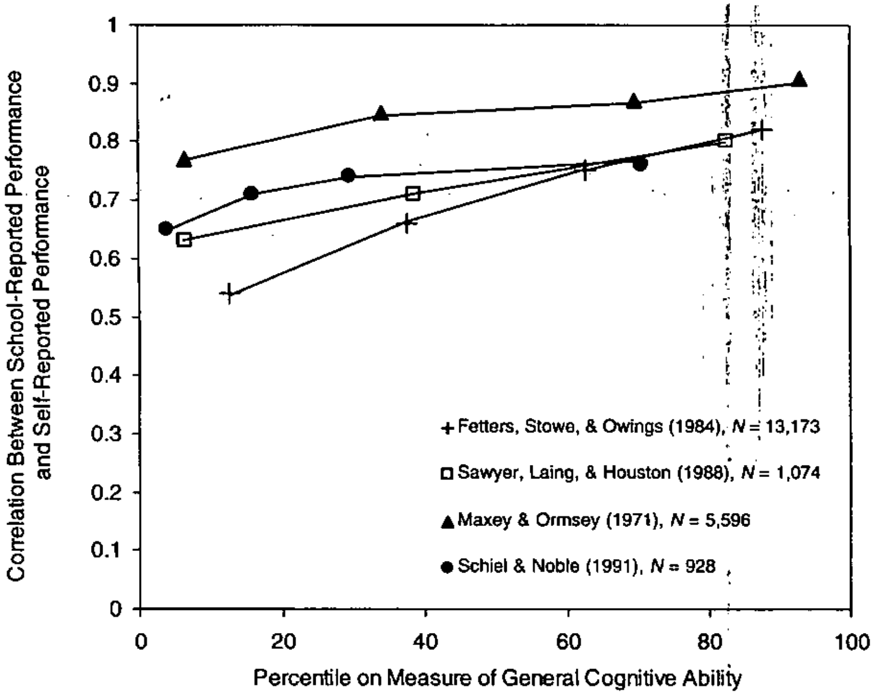

FIGURE 3. *Self-reported GPA reliability at various levels of general cognitive ability.*

Results

Results for the analyses of the accuracy of self-reported grades and ranks are presented in Table 1. The validity of GPA across all samples was relatively high ($N = 56{,}265$, $k = 29$, $r_{obs} = .84$), with college GPA being reported somewhat more accurately ($N = 12{,}089$, $k = 12$, $r_{obs} = .90$) than high school GPA ($N = 44{,}176$, $k = 17$, $r_{obs} = .82$). The validity of self-reported grades for individual subjects was lowest for art and music classes ($N = 1{,}627$, $k = 3$, $r_{obs} = .67$), and highest for social science classes ($N = 8{,}937$, $k = 8$, $r_{obs} = .85$). The validity of self-reported high school rank was slightly lower than the validity of self-reported high school GPA, both when high school rank was reported as a percentile ($N = 1{,}346$, $k = 3$, $r_{obs} = .76$) and when it was reported as a raw score ($N = 6{,}897$, $k = 3$, $r_{obs} = .77$).

There were no large differences in the validity of self-reported GPA of males ($N = 14{,}315$, $k = 7$, $r_{obs} = .79$) and females ($N = 13{,}179$, $k = 5$, $r_{obs} = .82$). The validity of self-reported GPA for White students ($N = 13{,}831$, $k = 3$, $r_{obs} = .80$) was higher than the validity of self-reported GPA for non-White students ($N = 5{,}544$, $k = 8$, $r_{obs} = .66$).

The validities of self-reported scores on standardized ability tests are presented in Table 2. The validities of self-reported SAT-Verbal ($N = 645$, $k = 5$, $r_{obs} = .74$), SAT-Mathematical ($N = 648$, $k = 5$, $r_{obs} = .80$), and SAT-Total scores ($N = 719$, $k = 6$, $r_{obs} = .82$) were comparable to the validities of self-reported high school GPA.

TABLE 1
Meta-analysis of correlations between self-reported grades or ranks and those obtained from school records

Academic variable	N	k	r_{obs}	SD_{obs}	SD_{rho}	90% credibility interval
Grade point average	56,265	29	.84	.07	.07	.70 to .94
College GPA	12,089	12	.90	.05	.05	.82 to .98
High school GPA	44,176	17	.82	.06	.05	.74 to .90
Math	9,492	9	.84	.05	.05	.76 to .92
Social science	8,937	8	.85	.02	.02	.82 to .88
Science	9,797	11	.82	.05	.05	.74 to .90
English	10,521	8	.84	.04	.04	.77 to .91
Physical sciences	756	2	.79	.02	.02	.76 to .82
Art/music	1,627	3	.67	.04	.03	.62 to .72
Foreign language	2,506	5	.84	.02	.02	.81 to .87
Natural science	5,494	5	.80	.05	.05	.72 to .88
High school rank (percentile)	1,346	3	.76	.02	.02	.73 to .79
High school rank (raw score)	6,897	3	.77	.01	.01	.75 to .79
GPA reported by men	14,315	7	.79	.05	.05	.71 to .87
GPA reported by women	13,179	5	.82	.04	.04	.75 to .89
GPA reported by Whites	13,831	3	.80	.08	.06	.70 to .90
GPA reported by non-Whites	5,544	8	.66	.04	.03	.61 to .71

Note. N = sample size, k = number of studies, r_{obs} = sample size weighted mean observed correlation, SD_{obs} = observed standard deviation of correlations, SD_{rho} = standard deviation of true score correlations.

The results for the meta-analyses of the standardized mean differences between self-reported GPA and actual GPA are reported in Table 3. Self-reported college GPA ($N = 6{,}507$, $k = 10$, $d = 1.38$) was over-reported to a far larger degree than high school GPA ($N = 4{,}566$, $k = 12$, $d = .32$), SAT-Verbal scores ($N = 594$, $k = 6$, $d = .33$), and SAT-Mathematical scores ($N = 578$, $k = 6$, $d = .12$).

TABLE 2
Meta-analysis of correlations between self-reported and actual SAT scores

Test	N	k	r_{obs}	SD_{obs}	SD_{rho}
SAT-Verbal	645	5	.74	.05	.03
SAT-Mathematical	648	5	.80	.09	.08
SAT-Total	719	6	.82	.09	.09

Note. N = sample size, k = number of studies, r_{obs} = sample size weighted mean observed correlation, SD_{obs} = observed standard deviation of correlations, SD_{rho} = standard deviation of true score correlations.

TABLE 3
Meta-analysis of standardized mean differences between self-reported GPAs and GPAs obtained from school records

Academic variable	*N*	*k*	*d*	SD_{obs}	SD_d
College GPA	6,507	10	1.38	.46	.44
High school GPA	4,566	12	0.32	.09	.00
SAT-Verbal	594	6	0.33	.16	.00
SAT-Mathematical	578	6	0.12	.04	.00

Note. N = sample size, k = number of studies, d = sample size weighted mean observed effect size, SD_{obs} = observed standard deviation of correlations, SD_d = standard deviation of true effect sizes.

The meta-analyses of the percentage of students who over-report and under-report their grades are presented in Table 4. The proportion of accurately reported grades was higher for high school GPA (N = 29,541, k = 10, % = 82.4) than for college GPA (N = 1,093, k = 8, % = 54.3). The incidence of over-reporting of grades was also higher for college GPA (N = 640, k = 3, % = 34.5) than for high school GPA (N = 26,481, k = 6, % = 12.3). Similarly, a higher percentage of college GPAs were under-reported (N = 640, k = 3, % = 8.8) than was the case for high school GPAs (N = 26,481, k = 6, % = 3.5).

The incidences of under-reported grades, accurately reported grades, and over-reported grades were similar for men and women, and for White and non-White students. Only 36.1% of SAT-Total scores (N = 292, k = 4) were reported accurately, with a far larger proportion of scores being over-reported (54.8%) than under-reported (12.1%).

Figure 2 displays the relationship between actual levels of school performance and the reliability of self-reported grades. The results clearly indicate that lower levels of school performance are associated with considerably lower levels of reliability for self-reported grades. That is, actual school performance moderates the reliability of self-reported grades. Figure 3 displays the relationship between students' levels of cognitive ability and the reliability of self-reported grades. Again, a moderating effect is observed, such that students with lower levels of cognitive ability (as measured by standardized admissions tests) tend to report their GPAs less reliably.

Discussion

Results suggest that self-reported grades are reasonably good reflections of actual grades for students with high ability and good grade point averages. However, self-reported grades are unlikely to represent accurately the actual scores of students with low GPAs and, to a lesser extent, low ability. Analyses that attempt to examine interactions or non-linearity across GPAs are not likely to yield accurate results when self-reported GPAs are employed. These findings may generalize to self-reports of other accomplishments.

High school GPA was, on average, somewhat less reliably reported (r_{obs} = .82, N = 44,176) than college GPA (r_{obs} = .90, N = 12,089). We suspect that some of this

TABLE 4
Meta-analysis of percentages of over-reporting, accurate reporting, and under-reporting of GPA, class rank, and SAT total scores

Academic variable	% Over	*N*	*k*	% Accurate	*N*	*k*	% Under	*N*	*k*
High school GPA	12.3	26,481	6	82.4	29,541	10	3.5	26,481	5
College GPA	34.5	640	3	54.3	1,093	8	8.8	640	3
GPA from previous quarter/semester	—	—	—	57.6	326	4	—	—	—
High school rank	—	—	—	55.0	540	4	—	—	—
GPA men	15.0	6,010	3	76.1	7,901	12	3.5	6,010	3
GPA women	11.5	5,971	2	81.5	7,826	8	2.6	5,971	2
GPA White	12.5	9,888	2	82.1	11,238	3	3.1	9,888	2
GPA non-White	12.8	1,220	7	80.6	1,420	8	2.2	1,220	7
SAT total scores	54.8	292	4	36.1	292	4	12.1	292	4

Note. % over = percentage of over-reporting; N = total sample size for all studies contributing to the estimate of percentages of over-reporting, accurate reporting, and under-reporting; k = number of studies; % accurate = percentage of accurate reporting; % under = percentage of under-reporting.

difference is due to the use of more complex GPA calculation methods that are often employed in some high schools. For example, it is not uncommon for high schools to employ an honors system in which honors classes receive bonus points. In this case, students may be confused about what grades or GPAs to report. When collecting GPAs from students, specific instructions may increase reliability.

Results also indicate that demographic variables do not substantially moderate the validity of self-reported grades. The only exception was for analyses examining minority students. The accuracy of their self-reported grade point averages appeared to be lower on average than those of non-minority students. Given the evidence that, on average, minority students receive lower grades than non-minority students (e.g., Morgan, 1990), these findings are consistent with the apparent moderating effect of actual school performance on the accuracy of self-reported grades. As with all moderator analyses on demographic variables, it is certainly possible that observed effects are due to unexamined variables that are associated with demographic variables. Therefore, the true effect could easily be due to socioeconomic or other experiential differences that happen to covary with race.

Subject area did moderate the validity of self-reported grades to a small extent. Lower correlations were found for the physical sciences and art and music classes. Our interpretation of these results is that correlations tend to be somewhat lower for disciplines for which there may be some confusion (e.g., chemistry, which can be thought of either as a physical or a natural science). The high school courses may also reflect respondent confusion (e.g., a respondent wondering if band participation is a music course). It is also likely that some music and art courses employ pass/fail grading, which may or may not be factored into a students' estimate of GPA. All of these factors can lead to disagreements between student self-reports and institutional records of GPA.

This study can help to identify situations where more faith can be placed in the results obtained with self-reported grades. The ideal situation would be to collect self-reported grades from college students who have done well in school and have high cognitive ability scores. These results also indicate that grades for particular subjects tend to be more reliable and suggest that researchers should provide specific information about what specific grades they do or do not want. The information could include specifying how specific topics should be categorized (e.g., all chemistry courses should be considered a physical science course).

Although systematic biases appear to influence (sometimes strongly) the validity of self-reported grades and grade point averages, it should be kept in mind that self-reported grades generally predict outcomes to a similar extent as actual grades. For example, self-reports of prior grade point averages are often excellent predictors of future grade point averages (Baird, 1976).

One solution is to treat the correlational results as simple reliability estimates. They could be used to correct correlation results from other studies and to adjust confidence intervals outward to reflect the reduced confidence in the estimates. Of course, this approach assumes that the difference between self and school-reported GPAs is purely a function of random error. The results presented in this study suggest that the relationship is a function of both random error and systematic biases, which are most common for students who have low GPAs. The upward bias in self-reported GPAs is likely to result in some indirect restriction of range and likely to introduce some non-GPA related variance.

Typically, we would expect that those with the lowest grades or scores would have the most to gain from misrepresenting their scores. However, all of the data examined in this study were for situations where the research participant could obtain no objective gain by misrepresenting self-reported grades. This suggests to us that intentionally inflated scores were provided for one of two reasons. Either the respondents did not believe what they were told by the parties requesting self-reported grades, or they felt that they would gain something by misrepresenting their GPA, such as protecting their pride or self-respect. Additional research is needed in this area.

Limitations and Suggestions for Future Research

Meta-analyses have received increased attention in education and psychology because of their utility for integrating literatures. Although quantitative integrations of the literature can be very powerful, they have a potential limitation. Research literatures are almost never random samples across people or situations. Thus the analyses may better reflect certain populations or situations limiting the generalizability of the findings. Although we obtained literature from a variety of sources and have several samples that approximate a good cross section of high school and college students, we cannot claim random samples or fully representative samples.

The nature of meta-analyses also tends to result in individual analyses that are less robust than would be ideal, because not all studies report complete moderator information. Although all analyses are based on hundreds of individuals, some are based on small numbers of studies. Additional research is needed to solidify the evidence for specific moderators. Personality traits beyond self-monitoring may prove to be important moderators. In the area of personality assessment, a number of traits, including integrity, conscientiousness, and neuroticism, appear to be associated with faking behavior (Mueller-Hanson, Heggestad, & Thornton, 2003; Snell, Sydell, & Lueke, 1999). These traits would be ideal places to begin examining personality traits that are predictors of inaccurate self-reports of grades.

Given the importance of self-reported grades to researchers, one particularly helpful area for future research would be the examination of what types of rating formats or data collection instructions will tend to yield the most accurate ratings. Because the type of instructions given appears to reduce score inflation on personality tests (Dwight & Donovan, 2003), similar procedures may also improve self-reported grades and grade point averages. A second practically important study for future research would be obtaining additional information about how the construct validity of self-reported grades declines over time. The studies examined here were from fairly short time frames, and some data suggests that, over time, self-reported grades become increasingly inaccurate until reaching an asymptote (Bahrick, Hall, & Dunlosky, 1993). Additional research is needed to more firmly establish the accuracy of self-reported grades over long periods of time.

Finally, as in all studies, the examination of moderators in a meta-analysis is constrained by the available data. The moderators examined here are associated with other variables that may, in fact, be the actual cause of the observed moderator effect. Ongoing research is needed to disentangle these interrelations and understand their causes.

The results of this meta-analysis and review clarify why there have been differing views on the utility of self-reported grades, test scores, and class ranks. The relationship between self-reported and school-reported grades is strong but far from unity. Furthermore, the evidence suggests that a non-trivial number of the errors appear to be systematic biases that are related to other individual differences. Because a goal of the field is to maximize the information value in the literature, we advocate using self-reported grades with caution and trusting their veracity only when the findings with self-reported grades mirror those obtained with actual grades. That is, we should treat self-reported results as replications when there is clear evidence that the findings agree with other research that used school-reported grades. Results based on self-reported grades can also be regarded with greater confidence when the nature of the sample and situation correspond to samples and situations found to yield more highly reliable information (e.g., students with stronger GPAs and college students). If clear discrepancies develop, then additional research will be needed. Overall, self-reported grades can be used, but with caution, and all research that employs self-reported grades must be evaluated with the findings reported here in mind.

Notes

We would like to thank Meghan Dalton, Elizabeth Varones, Kim Kosenga, Rebecca Wilson, Jennie Cordis, and Patrick Walsh for their assistance in gathering the articles summarized in this meta-analysis and David Klieger for his assistance with manuscript preparation.

Support for this project was provided by The College Board. Opinions expressed here are those of the authors and not necessarily those of The College Board.

Correspondence concerning this article should be addressed to Nathan R. Kuncel, Department of Psychology, University of Illinois, Champaign, IL 61820; e-mail *nkuncel@uiuc.edu*.

References

References marked with an asterisk are studies included in the meta-analyses.

*ACT. (1965). *Technical report for the ACT assessment program: Assessing students on the way to college.* Iowa City: ACT Publications.

Albrecht, D. D., Carpenter, D. S., & Sivo, S. A. (1994). The effect of college activities and grades on job placement potential. *National Association of Student Personnel Administrators Journal, 31,* 290–297.

*Armstrong, R. J., Jensen, J. A., Doyle, A. J., & Reynolds, C. H. (1976). The accuracy of the mean of six self-reported grades. *College and University, 21,* 183–192.

*Armstrong, R. J., Jensen, J. A., McCaffrey, R. F., & Reynolds, C. H. (1976). The accuracy of self-reported class rank. *Journal of the National Association of College Admissions Counselors, 21,* 37–42.

*Bachman, J. G. (1970). *Youth in transition: The impact of family background and intelligence in tenth-grade boys.* Ann Arbor: Braun-Brumfield.

*Bahrick, H. P., Hall, L. K., & Berger, S. A. (1996). Accuracy and distortion in memory for high school grades. *Psychological Science, 7,* 265–271.

Baird, L. L. (1976). Using self reports to predict student performance. *College Entrance Examination Research Monograph, 7,* 1–17.

*Bernard, L. C., & Walsh, R. P. (2002). Variations in a university subject pool as a function of earlier or later participation and self-report: A replication and extension. *Psychological Reports, 91,* 553–570.

*Cassady, J. C. (2001). Self-reported GPA and SAT: A methodological note. *Practical Assessment, Research, & Evaluation, 7*(12). Retrieved March 5, 2003, from http://:ericae.net/pare/getvn.asp?v=7&n=12

*Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt Brace Jovanovich College Publishers.

*Davidsen, O. M. (1963). *Reliability of self-reported high school grades.* Unpublished research report, American College Testing Program.

*Davidsen, O. M. (1963). *Survey of reliability of student self-reported high school grades.* (Available from the American College Testing Program, 500 ACT Drive, P.O. Box 168, Iowa City, IA 52243-0168)

*Dobbins, G. H., Farh, J., & Werbel, J. D. (1993). The influence of self-monitoring on inflation of grade-point averages for research and selection purposes. *Journal of Applied Social Psychology, 23,* 321–334.

*Dornbusch, S. M., Ritter, P. L., Leiderman, P. H., Roberts, D. F., & Fraleigh, M. J. (1987). The relation of parenting style to adolescent school performance. *Child Development, 58,* 1244–1257.

*Dunnette, M. D. (1952). Accuracy of students' reported honor point averages. *Journal of Applied Psychology, 26,* 20–22.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 6,* 1–23.

*Fetters, W. B., Stowe, P. S., & Owings, J. A. (1984). *Quality of responses of high school students to questionnaire items. High School and Beyond: A national longitudinal study for the 1980's.* Washington, DC: National Center for Education Statistics.

*Freeberg, N. E. (1988). *Analysis of the Revised Student Descriptive Questionnaire, Phase I: Accuracy of student reported information.* (College Board Report No. 88-5). New York: College Board Publications.

*Freeberg, N. E., Rock, D. A., & Pollack, J. (1989). *Analysis of the revised Student Descriptive Questionnaire: Phase II predictive validity of academic self-report.* New York: College Entrance Examination Board.

*Frucot, V. G., & Cook, G. L. (1994). Further research on the accuracy of students' self-reported grade point averages, SAT scores, and course grades. *Perceptual and Motor Skills, 79,* 743–746.

*Goldman, B. A., Flake, W. L., & Matheson, M. B. (1990). Accuracy of college students' perceptions of their SAT scores, high school and college grade point averages relative to their ability. *Perceptual and Motor Skills, 70,* 514.

*Hamilton, L. C. (1981). Sex differences in self-report errors: A note of caution. *Journal of Educational Measurement, 18,* 221–228.

*Hanna, G. S., Bligh, H. F., & Lenke, J. M. (1970). Student-reported grades as predictors. *Personnel and Guidance Journal, 48,* 465–468.

Humphreys, L. G., & Drasgow, F. (1989). Some comments on the relation between reliability and statistical power. *Applied Psychological Measurement, 13,* 419–425.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Thousand Oaks, CA: Sage Publications.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

*Jung, S. M., & Moore, J. C. (1970). The accuracy of self-reported high school grades. *College and University, 45,* 225–229.

*Kirk, B. A., & Sereda, L. (1969). Accuracy of self-reported college grade averages and characteristics of non and discrepant reporters. *Educational and Psychological Measurement, 29,* 147–155.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2004, August). *The validity of the GMAT: A meta-analysis.* Paper presented at the annual conference of the Academy of Management, New Orleans.

Kuncel, N. R., Crede, M., Thomas, L. L., Klieger, D. M., Seiler, S. A., & Woo, S. (in press). A meta-analysis of the Pharmacy College Admission Test (PCAT) and grade predictors of pharmacy student success. *American Journal of Pharmaceutical Education.*

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994a). Non-linearity of range restriction corrections in meta-analysis: Test of an improved procedure. *Journal of Applied Psychology, 79,* 425–438.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994b). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology, 79,* 978–986.

Linn, R. L., & Hastings, C. N. (1984). A meta-analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement, 21,* 245–259.

*Martin, C. L., & Nagap, D. H. (1989). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology, 74,* 72–80.

*Maxey, E. J., & Ormsby, V. J. (1971). *The accuracy of self-report information collected on the ACT Test Battery: High school grades and items of nonacademic achievement* (ACT Research Rep. No. 45). Iowa City, IA: American College Testing Program.

McCornack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology, 40,* 343–344.

Morgan, R. (1990). Analyses of predictive validity within student categorizations. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), *Predicting college grades* (pp. 225–238). Princeton, NJ: Educational Testing Service.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. III. (2003). Individual processes differences in impression management: An exploration of the psychological processes underlying faking. *Journal of Applied Psychology, 88,* 348–355.

*Nichols, R. C. (1966). Non-intellective predictors of achievement in college. *Educational and Psychological Measurement, 26,* 899–915.

*Nichols, R. C., & Holland, J. L. (1963). Prediction of the first year college performance of high aptitude students. *Psychological Monograph, 77,* 1–29.

*Paterson, D. G., & Thornburg, P. M. (1927). High school scholarship standing of freshman engineering students. *Journal of Engineering Education, 35,* 808–811.

*Payne, H., Harper, H., Quandt, B., Campbell, T., Bodenheimer, T., & White, L. (1995). Accuracy of college honors students' self-reported American Collegiate Test Scores. *Perceptual and Motor Skills, 81,* 64–66.

*Perry, J. D. (1940). The reliability of high school averages computed from students' estimates of their high school grades. *School and Society, 52,* 63–64.

*Peterson, R. E. (1968). *College student questionnaires technical manual.* Princeton, NJ: Institutional Research Program for Higher Education, Educational Testing Service.

Ramist, L. (1984). Predictive validity of the ATP tests. In T. F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 141–170). New York: College Entrance Examination Board.

*Richards, J. M., Holland, J. L., & Lutz, S. W. (1967). Prediction of student accomplishment in college. *Journal of Educational Psychology, 58,* 343–355.

Richards, J. M., & Lutz, S. W. (1967a). *Predicting student accomplishment in college from the ACT assessment* (ACT Research Report Series 21). Iowa City, IA: American College Testing Program.

*Richards, J. M., & Lutz, S. W. (1967b). *Predicting student accomplishment in college from the ACT assessment* (ACT Research Report Series 21). Iowa City, IA: American College Testing Program.

Roth, P. L., Be Vier, C. A., Switzer, F. S., & Schippman, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81,* 548–556.

Roth, P. L., & Clarke, R. L. (1998). Meta-analyzing the relation between grades and salary. *Journal of Vocational Behavior, 53,* 386–400.

*Sawyer, R., Laing, J., & Houston, M. (1988). *Accuracy of self-reported high school courses and grades of college-bound students* (ACT Research Report Series 88-1). Iowa City, IA: American College Testing Program.

*Schiel, J., & Noble, J. (1991). *Accuracy of self-reported course work and grade information of high school sophomores* (ACT Research Rep. No. 91-6). Iowa City, IA: American College Testing Program.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643–661.

*Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T.-Y. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88,* 979–988.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Toronto: Hogrefe & Huber.

*Shepperd, J. (1993). Student derogation of the Scholastic Aptitude Test: Biases in perceptions and presentations of College Board scores. *Basic and Applied Social Psychology, 14,* 455–473.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219–242.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30,* 526–537.

Snyder, M. (1987). *Public appearances, private realities: The psychology of self-monitoring.* New York: W. H. Freeman/Times Books/Henry Holt, & Co.

*Trice, A. D. (1990). Reliability of students' self-reports of scholastic aptitude scores: Data from juniors and seniors. *Perceptual and Motor Skills, 71,* 290.

*Trost, G., & Copony, R. (1976). Reliability of self-reported university examination grades. *Zeitschrift Fur Experimentelle und Angewandte Psychologie, 23,* 151–162.

*Valiga, M. J. (1986). *The accuracy of self-reported high school course and grade information* (ACT Research Report Series 87-1). Iowa City, IA: American College Testing Program.

*Walsh, W. B. (1967). Validity of self-report. *Journal of Counseling Psychology, 14,* 18–23.

*Walsh, W. B. (1968). Validity of self-report: Another look. *Journal of Counseling Psychology, 15,* 180–186.

*Walsh, W. B. (1969). Self-report under socially undesirable and distortion conditions. *Journal of Counseling Psychology, 16,* 569–574.

Whetzel, D. L., & McDaniel, M. A. (1988). Reliability of validity generalization data bases. *Psychological Reports, 63,* 131–134.

Willingham, W. W., & Breland, H. M. (1982). *Personal qualities and college admissions.* New York: College Entrance Examination Board.

*Windle, M. W., & Windle, R. C. (1996). Coping strategies, drinking motives, and stressful life events among middle adolescents: Associations with emotional and behavioral problems and with academic functioning. *Journal of Abnormal Psychology, 105,* 551–560.

Zakzanis, K. K. (1998). The reliability of meta-analytic review. *Psychological Reports, 83,* 215–222.

*Zimmerman, M. A., Caldwell, C. H., & Bernat, D. H. (2002). Discrepancy between self-report and school-record grade point average: Correlates with psychosocial outcomes among African American adolescents. *Journal of Applied Social Psychology, 32,* 86–109.

Authors

NATHAN R. KUNCEL is an Assistant Professor of Industrial and Organizational Psychology at the University of Illinois, Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820; e-mail *nkuncel@uiuc.edu.* He primarily studies the prediction and structure of academic and work performance. One of his career goals is to play matchmaker between methods, research, and theory from education and industrial/organizational psychology. After August 1, 2005, he will be at the University of Minnesota, Department of Psychology, 75 East River Road, Minneapolis, MN 55455.

MARCUS CREDÉ is a doctoral student in the Department of Psychology, University of Illinois, Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820; e-mail *crede@uiuc.edu.* His research interests include cognitive and noncognitive determinants of performance in both organizational and college settings.

LISA L. THOMAS is a doctoral student in industrial and organizational psychology, Department of Psychology, University of Illinois, Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820; e-mail *llthomas@uiuc.edu.* Her research interests include personnel selection, the prediction of performance in educational and work settings, personality measurement, and meta-analytic methods.