aps
ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
5(3) 233–242

DOI: 10.1177/1745691610369339
http://pps.sagepub.com
SAGE

Detecting and Correcting the Lies That Data Tell

Frank Schmidt
Department of Management, Tippie College of Business, University of Iowa, Iowa City

Abstract

Because of the way in which data are typically analyzed and interpreted, they frequently lie to researchers, leading to conclusions that are not only false but more complex than the underlying reality. The several examples of this presented in this article illustrate the possibility that although data may appear to indicate complex phenomena at the surface structure level, the phenomena may be quite simple at the deep structure level, suggesting the possibility of applying Occam's razor to achieve the scientific goal of parsimony. The approaches to data analysis described in this article may also lead to a solution to the serious problem of construct proliferation in psychology by demonstrating that many constructs are redundant with other existing constructs. The major obstacles to these outcomes are researchers' continued reliance on the use of statistical significance testing in data analysis and interpretation and the failure to correct for the distorting effects of sampling error, measurement error, and other artifacts. Some of these problems have been addressed by the now widespread use of meta-analysis, but examination of the meta-analyses appearing in *Psychological Bulletin* from 1978 to 2006 shows that most employ a statistically inappropriate model for meta-analysis (the fixed effects model) and that 90% do not correct for the biasing effects of measurement error. Hence, there is still a long way to go in the improvement of data analysis and interpretation methods.

Keywords

data analysis, meta-analysis, measurement error, theory development, cumulative knowledge

The data that psychologists base their research conclusions on are often deceptive. My interest in probing the hidden meaning of data began with my dissertation 40 years ago (Schmidt, 1970). That dissertation showed that supposedly statistically optimal regression weights often produce less accurate prediction in new samples and in the population than do simple equal (unit) weights (Schmidt, 1971, 1972). This was very surprising to many. Other studies focusing on the hidden meaning of data followed (for example, Schmidt, Berner, & Hunter, 1973; Schmidt, Hunter, & Urry, 1976). This interest led to my work with the late Jack Hunter in developing meta-analysis methods aimed at revealing the true meaning of research literatures consisting of apparently conflicting studies (Schmidt & Hunter, 1977). Over the years, these methods have been presented and tested in numerous journal articles and in three books (Hunter, Schmidt, & Jackson, 1982; Hunter & Schmidt, 1990b, 2004).

These methods have been applied thousands of times and have revealed that the conflicts in the literature are often more apparent than real. In the process, this work has revealed the ways in which research data, when taken either at face value or interpreted using significance tests, lie to researchers, leading to false conclusions. In considering the ways in which research data are typically interpreted, I became convinced that there is a strong cult of naive and overconfident empiricism in psychology and the social sciences with an excessive faith in data as the direct source of scientific truth and an inadequate appreciation of how misleading data can be. I concluded that the commonly held belief that research progress requires only that we "let the data speak" is sadly erroneous. If data are allowed to speak for themselves, they will typically lie to you. This article presents several examples of this.

The goal of psychology, as it is for other sciences, is cumulative knowledge. To develop theories that constitute such knowledge, we must know the relations between variables and constructs, because such relations are the building blocks of theory. Ironically, our most frequently used data analysis and

Corresponding Author:
Frank Schmidt, Department of Management, Tippie, College of Business, University of Iowa, Iowa City, IA 52242
E-mail: frank-schmidt@uiowa.edu

Table 1. Raw Data for Decision-Making Test (16 Studies)

Study	*r*	*N*
1	.20*	203
2	.11	214
3	.02	225
4	.16	38
5	.34	34
6	.13	94
7	.31*	41
8	.25*	323
9	.32*	142
10	.39	24
11	.22*	534
12	.17	30
13	.20*	223
14	.24*	226
15	.24*	101
16	.19	46

* $p < .05$

interpretation methods retard or prevent achievement of this goal.

An Example of Lying Data

I would now like to present an example of this using a real data set. Table 1 shows data from 16 real studies done in 16 different organizations relating scores on a test of decision making to supervisory ratings of job performance in various midlevel jobs. Figure 1 shows the histogram of these correlations, which makes clear the great variation in outcomes. Figure 2 shows the approximate long-run expected distribution of these correlations—that is, the distribution expected if the number of studies were much larger (for example, 1,600 studies instead of 16).

Returning now to Table 1, we can see that there is great variability in both sample sizes and correlations. The sample sizes range from 24 to 534. The correlations range from .02 to .30, a ratio of 15 to 1. If we take these data at face value, we would conclude that there is tremendous variation in magnitude but that all relationships are positive. However, that is not the most common mode of data interpretation. Instead, significance tests are used. After we apply significance tests, we see that eight studies (half) are nonsignificant at the .05 level (two-tailed) and half are significant. Results are maximally conflicting. One common conclusion is that there is no relationship in eight of the organizations and a positive relationship in the other eight organizations. This is the moderator (interaction) interpretation—which leads to the conclusion that there is a major moderator separating the 16 organizations into two different types. This conclusion is based on the typical practice of concluding that there is no relationship when a result is not statistically significant. Another common conclusion is that based on the majority vote rule. This common rule leads to the conclusion that there is no relationship between these two variables because a majority of the relationships are nonsignificant. We will see shortly that the interpretations using significance

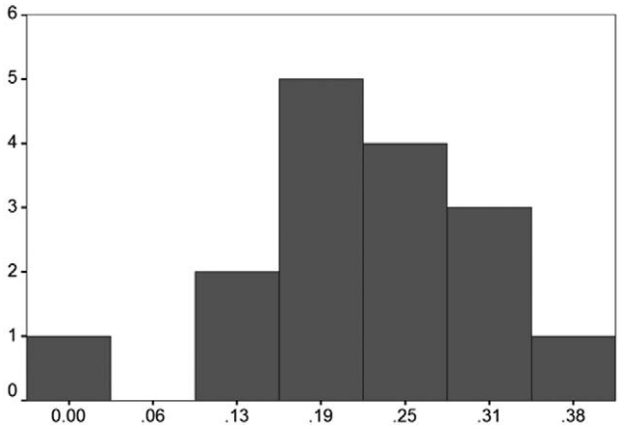

Fig. 1. Histogram of raw data for decision-making test.

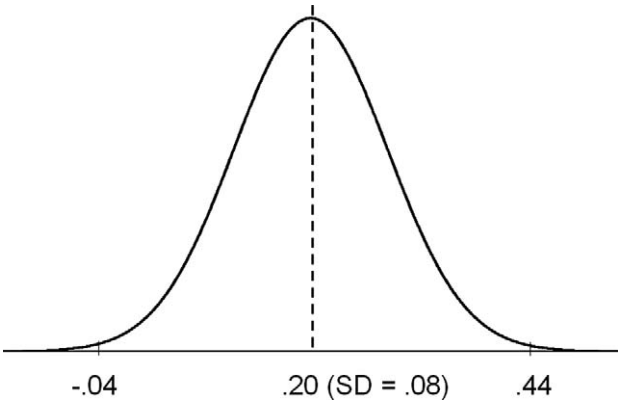

Fig. 2. Approximate expected distribution of raw data for decision-making test.

tests are more erroneous than the naive interpretation that accepts the correlations at face value and does not employ significance tests.

Figure 3 shows the distribution when these correlations are corrected for sampling error. Sampling error is the random departure of statistical estimates computed on samples from values in the population (the values of interest). Sampling errors vary randomly around zero, and the smaller the sample, the more widely they vary. You can see that almost all the variation (98%) was due to simple sampling error! The square root of .98 is .99; this is the correlation between individual study sampling errors and observed values of the correlations! The remaining variance is very, very small. The *SD* is only .01, and the variance is only .0001. This illustrates the important fact that researchers routinely underestimate the impact of sampling error on their data. How many researchers realize that sampling error can produce such wide variation in results? How many researchers are aware the observed estimates in their data could be highly correlated with the sampling errors in those data? (Note that even if only 70% of the variance were explained by sampling error, this correlation would still be high: .84, the square root of .70.) Controlling for only sampling error shows that these studies are quite consistent with each other.

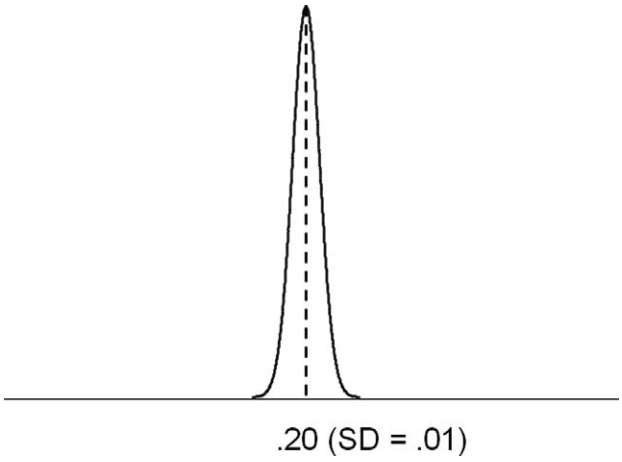

Fig. 3. Distribution of data corrected for sampling error only.

All indicate a relationship of .20, or very nearly .20. For an even more dramatic example of this process, see Schmidt, Ocasio, Hillery, and Hunter (1985).

The methods used here to correct for sampling error are from the Hunter and Schmidt (2004) meta-analysis book (and the two books cited earlier). They are implemented using the Schmidt and Le (2004) Windows-based meta-analysis program package. Essentially, one computes the amount of variance expected from sampling error, using known formulas for sampling error variance, and subtracts this value from the observed variance of the correlations.

Now we shall consider measurement error. All of these correlations are biased downward by measurement error in both measures. Measurement error, like sampling error, exists in all data. There are no exceptions—because there are no perfectly reliable measures. Suppose we want to estimate the construct level relationships—the usual value of interest in research developing and testing theories (Hunter & Schmidt, 2004; Schmidt & Hunter, 1996, 1999). To do this, we use the reliabilities of each measure to correct each correlation. The reliability of the decision-making test is .80 and the (interrater) reliability of the ratings of job performance (by one rater) is .50. These reliability values are used with each correlation in the standard formula for correcting for the biases created by measurement error (the classical disattenuation formula). The reliability values are the same in all the studies because the same measures were used in all organizations in this consortium study.

Table 2 shows the corrected *r*s and their significance levels. Note that the significance levels are not affected by the correction. This is because the correction increases the standard error of each correlation. The downward bias has been removed, so the correlations are all larger but still half are nonsignificant. Figure 4 shows the histogram of these values. You can again see the great variation. The correlations range from .03 to .62, a ratio of 20 to 1! Figure 5 shows the approximate long run expected distribution of the corrected correlations. Note the larger mean (.32) and larger *SD* (.12). The variability is even larger now than before. This is because the correction for measurement error increases sampling error variance. Figure 6 shows the results of the full meta-analysis, correcting for both sampling error and measurement error. Almost all variation is

Table 2. Data From Table 1 Corrected for Measurement Error

Study	r_c	*N*
1	.32*	203
2	.49	214
3	.52	225
4	.25	38
5	.54	34
6	.21	94
7	.49*	41
8	.39*	323
9	.51*	142
10	.62	24
11	.35*	534
12	.27	30
13	.32*	223
14	.39*	226
15	.39*	101
16	.30	46

* $p < .05$

Note: Table 2 contains two errors. The 2nd correlation should be .17, not .49; and the third correlation should be .03, not .52

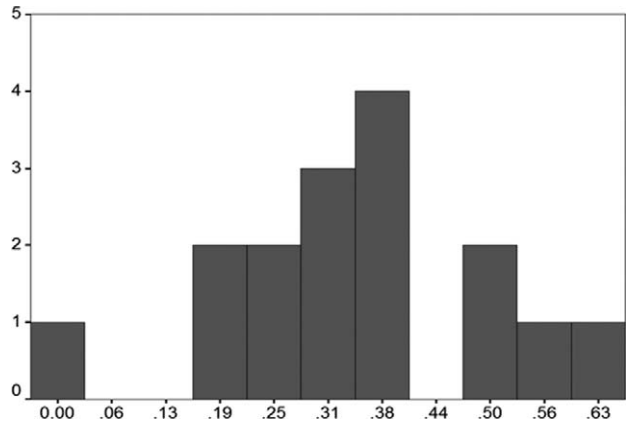

Fig. 4. Histogram of data corrected for measurement error only.

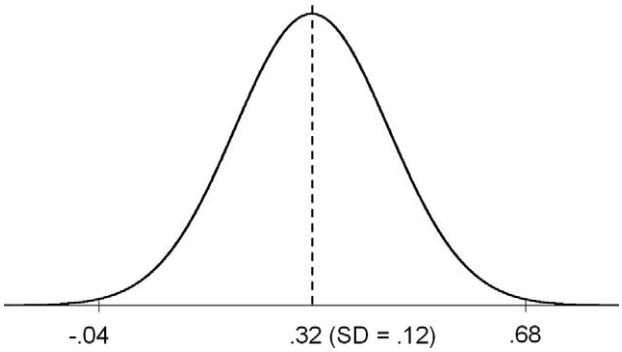

Fig. 5. Approximate expected distribution of data corrected for measurement error only.

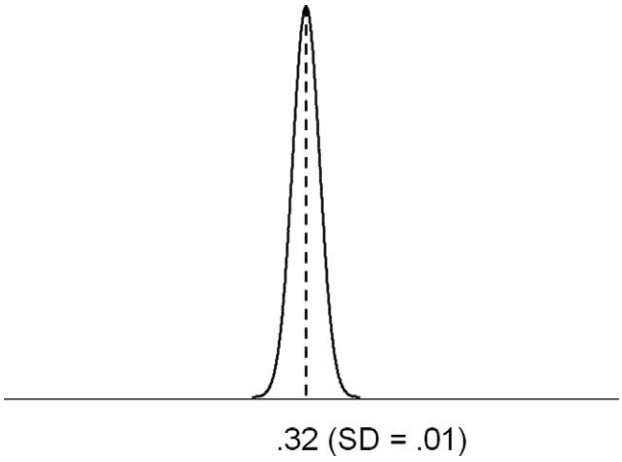

Fig. 6. Distribution of data corrected for both sampling error and measurement error.

now accounted for. In conclusion, there is basically a single value (.32) because there is nearly no variation.

Note that this interpretation is not only more accurate, it is also more parsimonious. That is, it is an example of application of Occam's razor in data interpretation. It is an example of simplicity underlying apparent complexity. The surface structure of the data is quite complex, but the deep structure is quite simple.

Contrast this conclusion with the earlier interpretations produced when we allowed the data to lie to us. One interpretation using significance tests was that there is a major moderator operating creating two different types of organization. The other interpretation using significance tests was that there is no relation between these two variables. The naive approach, taking the data at face value and not using significant tests, was that there was great variability in results but that all relations were positive. Note that this naive interpretation is less wrong than the interpretations using significance tests: The relationships are in fact all nonzero and positive. This example illustrates how data can and do lie. Again, remember that all data—without exception—are distorted by both sampling error and measurement error.

In this example, the same measures were used in all studies, so the reliabilities were the same for all studies. Usually this is not the case. The reliabilities vary, and this variation creates additional variance beyond sampling error. Correcting for measurement error then produces additional reductions in variance beyond that produced by correcting for sampling error. This did not happen here, but when it does, the Hunter–Schmidt method of meta-analysis corrects for this additional artifactual variation.

Other Examples of Lying Data

Now let's consider other examples. Ones (2008) studied expatriates—American executives and managers sent overseas by American multinational companies. She was interested in whether different personality traits predict success in different cultures. She found that any given personality trait was significantly correlated with expatriate success in some countries but not in others. This appeared to indicate that different personality traits are important in different countries—something that companies would need to consider when deciding which manager to send to which country. This would be the common interpretation of such findings. However, an analysis of the sort I illustrated showed that virtually all the variability in validity across countries was explained by sampling error! The apparent effect of cultural differences was a data illusion. Again, these methods reveal a pattern of simplicity underlying apparent complexity.

A similar situation of lying data was found in a recent dissertation by one of my PhD students (Oh, 2009). Oh focused on the criterion-related validity of the Big Five personality traits in East Asian countries (Korea, China, Japan, Taiwan, and Singapore). He hypothesized that because the cultures of these countries are different (and more different from each other than the cultural differences among European countries), they will show different patterns of validity for personality traits in predicting job performance. For each country, he meta-analyzed studies on each personality trait, and it appeared that there were indeed substantial mean differences in validity across countries. However, he recognized that each meta-analytic mean still contained some sampling error (second order sampling error; Hunter & Schmidt, 2004), and so he conducted a second order meta-analysis—that is, a meta-analysis of the mean values across countries. He found that second order sampling error accounted for all the between-country validity variance for four of the five personality traits. For the fifth trait (Conscientiousness), most of the variance (59%) was accounted for by second-order sampling error. Again, the deep structure of the data was found to be quite simple despite the fact that the surface structure gave the appearance of considerable complexity.

Now consider a fourth example—this one in traffic safety engineering (Hauer, 2004). A number of studies had been done to see whether highway shoulders, which allow drivers to pull over to the side of the road if they need to stop for any reason, reduce accidents and deaths. These small sample studies found no significant relationships, so traffic safety experts concluded that putting shoulders along roads did not reduce accidents or deaths. As a result, far fewer shoulders were built in most states, saving construction money. Hauer's (2004) meta-analysis of these studies showed clearly that shoulders reduced both accidents and traffic deaths. In this example, people died because of reliance on statistical significance tests in interpreting study data! Hauer found the same pattern of results for the 1987 and 1995 increases in speed limits on interstate highways. (For a detailed analysis of this problem in safety research, see Hauer, 1983.)

Other Considerations

The examples presented above show how meta-analysis can be used to precisely calibrate relationships of theoretical and practical interest. This procedure can also be used to create a matrix of relationships among several variables (constructs) and such a matrix can then be used in path analysis to test causal models or theories (Becker, 2009; Hunter & Schmidt, 2004; Schmidt,

Hunter, & Outerbridge, 1986). For example, suppose you have four variables in your causal model; this means there are six correlations among these variables. Separate meta-analyses can be used to estimate each of these correlations. The causal model can then be tested using path analysis, even though no one primary study may have estimated all of these six relationships. Note that this analysis is actually akin to structural equation modeling, as all correlations have been corrected for measurement error (and also for range restriction, if appropriate). This process is increasingly common in the literature today and is another important contribution of meta-analysis to development of cumulative knowledge (Becker, 2009).

The data in these examples are correlational. But the same principles apply to experimental data. The only difference is that the focal statistic is then usually the *d* value—the difference between two groups in standard deviation units. These groups can be an experimental group and a control group or any two groups (e.g., men and women). The same model of data analysis applies. The *d* value statistic is subject to even more sampling error than the correlation and is also biased downward by measurement error, so both corrections are again needed (Hunter & Schmidt, 2004). In an earlier article (Schmidt, 1992), I presented an example based on data from experiments.

Again, the program used to analyze these data is the Schmidt and Le (2004) program. This program first corrects each correlation for measurement error and then performs the meta-analysis on the corrected correlations. Space does not permit a discussion of the technical details of how this is done, but these are covered in detail in the Hunter and Schmidt (2004) meta-analysis book. These methods are the only meta-analysis methods that take into account both sampling error and measurement error, both of which are present in all data (Rothstein, 2003). These methods correct simultaneously for data distortions caused by both kinds of error. Hedges (2009) has noted the importance of making both kinds of corrections.

As just illustrated, the appearance of moderators can be illusory. However, in some cases there are in fact real moderators (interactions). For example, it has been shown in several comprehensive and independent meta-analyses of U.S. and European data that the information processing complexity of jobs moderates the size of the validity of intelligence tests in predicting job performance (e.g., see Hunter & Hunter, 1984; Salgado, Anderson, Moscoso, Bertua, & de Fruyt, 2003). The most current estimates indicate that this validity increases from .39 to .73 as jobs go from the simplest to the most complex (Hunter, Schmidt, & Le, 2006). Considerable evidence is required to confirm such a moderator, and this evidence has been presented in this case, unlike the cases presented in my examples. However, researches often report ''moderators'' that are not really moderators. An example is the proposition that the type of employment test (personality tests vs. intelligence tests) moderates predictive validity for job performance. As discussed later, it is true that the validity of intelligence tests is much higher than that of personality tests. But this is a pseudomoderator created by comparing apples and oranges: Intelligence tests and personality tests are completely different animals.

Measurement Error and Construct Redundancy

One of the major problems in psychology is construct proliferation. Researchers frequently postulate new constructs that are questionably different from existing constructs, a situation contrary to the canon of parsimony. For example, is job involvement really different from job satisfaction? Proper corrections for measurement error are now making another contribution: They are showing that some constructs are probably completely redundant at the empirical level. For example, our research has shown that measures of job satisfaction and organizational commitment correlate nearly 1.00 when each measure is appropriately corrected for measurement error. (Le, Schmidt, Harter, & Lauver, in press; see also Le, Schmidt, & Putka, 2009). These constructs are conceptually distinct, but not empirically distinct. They are conceptually distinct in that job satisfaction refers to one's reactions to his or her specific job, whereas organizational commitment refers to one's evaluation of the organization in which one works. Apparently, respondents do not make this distinction between these constructs, as researchers do. These findings are made possible by newly developed more accurate methods of correcting for measurement error (Le et al., 2009, in press). Findings like this have important implications for parsimony and theory construction in science. Again, the picture that emerges is one of simplicity underlying apparent complexity. The surface structure appears complex, but the deep structure is quite simple. And as scientists, we are interested in deep structure, not surface structure. Our goal is to ''carve nature at its joints.''

In my initial example, the most erroneous interpretations were those based on statistical significance tests, and this is generally the case. But, as demonstrated in a number of publications (e.g., Carver, 1978; Cohen, 1994; Loftus, 1996; Oakes, 1986; Schmidt, 1996; Schmidt & Hunter, 1997), researchers appear to be virtually addicted to significance testing, and they hold many false beliefs about significance tests:

1. ''If my finding is significant, I know it is a reliable finding and the probability of replication is about .95 (1 minus the *p* value).'' This is false. Statistical significance has no bearing on replication probability. Actually, the probability of replication is usually around .50, which is the typical level of statistical power in most research literatures (Cohen, 1962, 1988; Schmidt & Hunter, 1997; Sedlmeier & Gigerenzer, 1989).
2. ''The *p* value is an index of the importance or size of a relationship.'' Again, this is false. The *p* value is a function mostly of sample size. It provides no measure of the size or importance of the relationship.
3. ''If a relationship is not significant, it is probably just due to chance and the real relationship is zero.'' Also false. Actually, most nonsignificant findings are due to low

statistical power to detect relationships that do exist and are often important. As shown by Lipsey and Wilson's (1993) review of hundreds of meta-analyses, nonzero relationships are almost always the case. In their review, 300 of 302 meta-analyses (99%) showed nonzero relationships.

4. "Significance tests are necessary if we are to test hypotheses, and hypothesis testing is central to scientific research"—false. The physical sciences (e.g., physics and chemistry) routinely test hypotheses and do not use significance tests. In fact, most physical scientists view the use of significance tests as indicative of a pseudoscience (Schmidt & Hunter, 1997). The physical sciences use point estimates (effect sizes) and confidence intervals (CIs).
5. "Significance tests are essential because they ensure objectivity, which is critical in science." This is false. CIs are just as objective as significance tests, and they provide far more information (Borenstein, 1994; Cumming & Finch, 2005).
6. "The problem is only the misuse of significance tests, not the tests themselves." This too is false. Even when not misinterpreted, significance tests retard the development of cumulative knowledge, whereas point estimates, effect size estimates, and CIs promote cumulative knowledge (e.g., see Schmidt & Hunter, 1997, and Thompson, 2002).

Some steps have been taken to address the problems created by overreliance on significance testing. The 1999 APA Task Force Report on significance testing (Wilkinson and The APA Task Force on Statistical Inference, 1999) states that researchers should report effect size estimates and CIs. And both the 5th and 6th editions of the APA Publication Manual state that it is "almost always" necessary for studies to report effect size estimates and CIs (American Psychological Association, 2001, 2009). The reporting standards of the American Educational Research Association (2006) have an even stronger requirement of this sort. Also, at least 24 research journals now require the reporting of effect sizes as a matter of journal policy (Thompson, 2007). However, as is apparent from perusing most psychology research journals, progress in this area still has a long way to go.

The problems created by use of significant tests in interpreting literatures are illustrated in the example I presented earlier in this article. These problems stem from use of statistical significance testing in individual studies; this practice should be replaced by point estimates of effect sizes and CIs. CIs typically show that the literature is not really conflicting, because they overlap across almost all studies (Schmidt, 1992).

Data Distortions Beyond Sampling and Measurement Error

My initial example illustrates the two artifacts that are always present in any literature. But there are others that are often, but not always, present, such as data errors, range restriction, dichotomization of measures, and imperfect construct validity. Data errors—typos, coding errors, transcription errors, etc.—have been shown to be very prevalent (Hunter & Schmidt, 2004, pp. 53–54). This is a nonsystematic source of variability, like sampling errors. Unless they result in extreme or impossible outliers, data errors are hard to identify and therefore difficult or impossible to correct.

Unlike data errors and sampling errors, range restriction is a systematic artifact. Range restriction reduces the mean correlation. Also, variation in range restriction across studies increases the between-study variability of study correlations. Differences across studies in variability of measures can be produced by direct or indirect range restriction (DRR and IRR). DRR is produced by direct truncation on the independent variable and on only that variable. For example, range restriction would be direct if college admission were based only on one test score, with every applicant above the cut score admitted and everyone else rejected. This is quite rare, because multiple items of information are almost always used in decision making. Most range restriction is indirect. For example, self-selection into psychology lab studies can result in IRR on study variables. Range restriction is correctable in a meta-analysis, and my research team has recently developed a new procedure for correcting for IRR that can be applied when older procedures are not feasible (Hunter et al., 2006). This procedure has been demonstrated via Monte Carlo simulation studies to be quite accurate (Le & Schmidt, 2006). Application of this new correction method has shown that general intelligence is considerably more important in job performance than previously thought. The correlation for the most common job group in the economy is about .65. Previous estimates of this correlation, based on corrections for DRR when in fact IRR existed in the data, have been about .50 (Hunter et al., 2006; Schmidt, Oh, & Le, 2006), so the new estimate is about 30% larger than the older one. This is a substantial difference.

Application of this new method has also changed estimates of the relative importance of personality and intelligence as determinants of job performance, showing that intelligence is nearly three times more important than the personality trait of Conscientiousness and five times more important than the personality trait of emotional stability (Schmidt, Oh, & Shaffer, 2008). (Personality measures were self-report scales of the Big Five personality traits.) A recent application of this IRR correction method has also shown that the Graduate Management Aptitude Test (GMAT) is more valid than previously thought (Oh, Schmidt, Shaffer, & Le, 2008). These are examples of how meta-analysis methods are continuing to evolve and develop, even 35 years after their introduction in the mid-1970s (Schmidt, 2008).

Another artifact is caused by dichotomization. Researchers often dichotomize continuous measures into "high vs. low" groups. This practice not only loses information but also lowers correlations and creates more variability in findings across studies (Cohen, 1983; Hunter & Schmidt, 1990a, 2004; MacCallum, Zhang, Preacher, & Rucker, 2002). The distorting effects of dichotomization are correctable in a meta-analysis.

The final additional artifact I want to mention is imperfect construct validity in measures. Even after correction for measurement error, the measure may correlate less than perfectly with the desired construct (Schmidt, Le, & Oh, 2009). This is especially true when proxy measures are used (for example, use

of education as a proxy for general mental ability). Degree of construct validity may vary across studies, causing between-study variability and typically lowering the mean. Correction for this requires special information, is complicated, and is often not possible (Hunter & Schmidt, 2004).

When it is not possible to correct for one or more of these artifacts, the researcher should take this fact into account when interpreting and reporting the meta-analysis results. The meta-analyst should clearly state that the variance left unaccounted for could be due to these uncorrected artifacts. This can help prevent erroneous conclusions that moderators exist when they do not.

So it is clear that the injunction to "just let the data speak" is very naive and deceptive. Data often look you in the eye and lie to you—without even blinking—and this is especially true when interpretation of data is based on significance testing, as my initial example illustrates. So what is the solution? In individual studies, researchers should use point estimates and CIs, not significance tests (Loftus, 1996; Schmidt, 1992; Thompson, 2002). As noted, there has been some movement in this direction. Also, researchers should correct for measurement error, dichotomization, and any other distorting effects that can be corrected in a single study. There has been less improvement in this respect, but there has been some. In the integration of literatures and reviews, researchers should use meta-analysis. Note that even if the first step is not realized, the second can still undo the damage done by erroneous data interpretation in individual studies, but only if the individual studies include the information needed to compute the effect sizes and make the needed corrections—not all do.

Random and Fixed Models in Meta-Analysis

Most readers are aware that meta-analysis is widely used today and that conclusions about cumulative knowledge presented in textbooks and elsewhere are increasingly based on meta-analysis results (Hunter & Schmidt, 2004, pp. 26–27). So isn't the problem solved? Actually, there is still a serious problem, because an inappropriate statistical meta-analysis model is very frequently used in the literature. In most research areas, some variability is usually left after correcting for artifacts. Only a random effects (RE) meta-analysis model can reveal whether this is the case or not. The example I presented is based on an RE model. Fixed effects (FE) meta-analysis models assume a priori that there is only a single population parameter underlying all studies. That is, FE models assume that all variation across studies is due to solely to sampling error and that therefore none of the variation is due to real differences between studies in underlying parameters.[1] This a priori assumption is highly questionable in most cases. RE models, by contrast, treat this assumption as an hypothesis and test it—allowing the researcher to see whether or not all variance is accounted for by sampling error and other artifacts.

The premier review journal in U.S. psychology is *Psychological Bulletin*, and the majority of reviews in that journal today are meta-analyses. In fact, narrative reviews are often returned to authors with instructions to perform meta-analyses (Hunter & Schmidt, 2004, pp. 26–27). My colleagues and I recently examined the meta-analyses in this journal (Schmidt, Oh, & Hayes, 2009) and found that a total of 199 meta-analyses were published in *Psychological Bulletin* between 1978 and 2006. Of the 169 that could be classified as either FE or RE models, 79% (129) used only FE models. Figure 7 shows these findings.

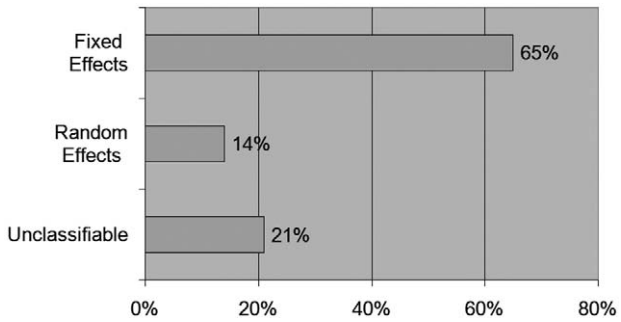

Fig. 7. Fixed versus random meta-analysis models in *Psychological Bulletin*, 1978–2006.

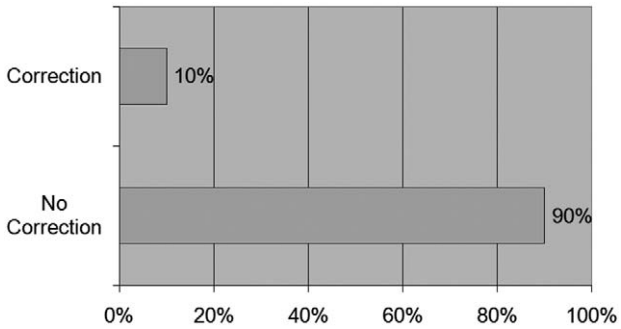

Fig. 8. Frequency of correction for measurement error in meta-analyses in *Psychological Bulletin*, 1978–2006.

A reanalysis of data from five of these FE meta-analysis studies (containing a total of 68 different meta-analyses) showed they seriously underestimated the width of the CIs they reported by an average of 52%. That is, the CIs were only half as wide as their real width, a gross overestimation of the precision of the findings. On average, the CIs reported as 95% CIs were actually 55% CIs (Schmidt, Oh, & Hayes, 2009).

What about corrections for measurement error? We found that 180 of the 199 published meta-analyses (90%) did not correct for measurement error—which, as noted earlier, is always present! Nor did they correct for the other artifacts I discussed. Figure 8 shows these findings.

The combined result of these two errors is mean values that are too low and CIs that are too narrow. Effect sizes are biased downward but the precision with which they are estimated is very much overstated! These are not mere technical differences. These are large differences that have important implications for conclusions, theory development, and practical value of applied procedures (Schmidt, Oh, & Hayes, 2009). Based on our examination of other psychology journals, we believe the situation in *Psychological Bulletin* is typical of much of the

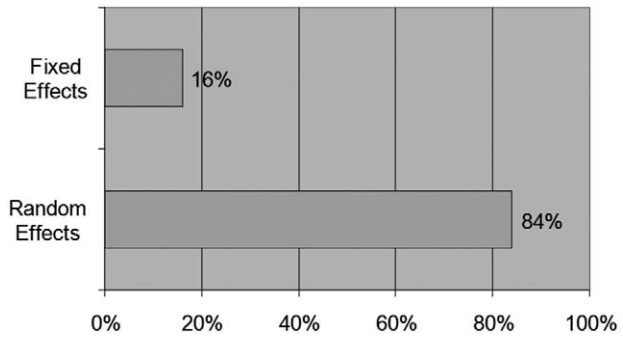

Fig. 9. Fixed versus random meta-analysis models in industrial/organizational psychology literature.

psychological literature. The only literature that seems to be an exception is the industrial/organizational (I/O) psychology literature—which is a large literature but still a small fraction of the overall psychology literature. Based on the content of the two top journals in the I/O field, only 16% of the meta-analyses use the FE model and 84% use the RE model, as shown in Figure 9. The figures are the same for corrections for measurement error: 84% of meta-analysis in the I/O literature do correct and only 16% do not.

So the problem of accurate analysis and integration of research literatures in psychology has not yet been fully solved. The improvements we have seen to date leave much to be done. On a positive note, however, the methods described above that correct not only for sampling error but also for measurement error, range restriction, and other artifacts have been applied many times in the literature. Some illustrative examples are listed in Table 3.

Table 3. Examples of Research Topics With Complete Meta-Analyses

Area of study	Topic
Organizational psychology	Relations of personality to leadership perceptions and effectiveness; role of trust in leadership; goal commitment, goal difficulty, and job performance; job and life satisfaction; organizational commitment and workplace outcomes
Personnel psychology	General mental ability and job performance in the United States and Europe, personality and job performance in the United States and Europe; effects of feedback intervention on performance; accuracy of self-ratings of ability and skill
Occupational/health psychology	Work–family conflict and life satisfaction, job and life satisfaction, correlates of role conflict and role ambiguity, gender differences in occupational stress, effectiveness of smoking cessation methods, back pain and absence from work
Individual differences/ differential psychology	Affective underpinnings of job perceptions and attitudes; personality and subjective well-being; The Big Five personality and Holland's vocational interest; relations among intelligence, personality, and interests; the role of person versus situation in life satisfaction; height, self-esteem, and career success
Education/vocational psychology	Effects of psychosocial and study skill factors on college outcomes, effects of learning goal orientation, the validity of the Graduate Record Examinations, career benefits of mentoring

Conclusion

I hope this article makes clear how getting the truth out of data has turned out to be a much more complicated and challenging process than we have traditionally thought. But this process is critical to the attainment of cumulative knowledge in psychology and therefore deserves our close attention and best efforts to get it right. It is not a matter that can be taken lightly. We cannot have a successful science if we let our data lie to us. To attain cumulative knowledge, we must detect and correct those lies. If we do this, we can successfully apply Occam's razor and uphold the important principle of scientific parsimony. We can discover the simplicity at the deep structure level that underlies the apparent and confusing complexity at the surface structure level.

Note

1. This assumption applies in the FE model in the usual case in which the findings are generalized beyond the study set included in the meta-analysis to the wider set of real and potential studies. This is almost always the case, because the goal in science is generalizable knowledge. If the meta-analysis is used only to describe the specific set of studies at hand, with no broader conclusions being drawn, then the FE model does not assume a single underlying population value. However, such an application is not very useful scientifically. Another relevant fact is that FE models in the literature allow only for artifactual variation produced by sampling error. They do not recognize variation produced by other artifacts (such as range restriction or measurement error). For a detailed discussion of these issues, see Schmidt, Oh, and Hayes (2009).

Acknowledgments

This article is based on the author's James McKeen Cattell Award address, presented at the 2008 convention of the Association for Psychological Science, Chicago, IL, on May 24. I would like to thank In-Sue Oh and Huy Le for their comments or an earlier draft and for other assistance on this article.

Declaration of Conflicting Interests

The author declared that he had no conflicts of interest with respect to his authorship or the publication of this article.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Becker, M.J. (2009). Model-based meta-analysis. In H. Cooper, L.V. Hedges, & J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 377–395). New York: Russell Sage Foundation.

Borenstein, M. (1994). The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials*, *15*, 411–428.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round *($p < .05$)*. *American Psychologist*, *49*, 997–1003.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170–180.

Hauer, E. (1983). Reflections on methods of statistical inference in research on the effects of safety countermeasures. *Accident Analysis and Prevention*, *15*, 275–285.

Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis and Prevention*, *36*, 495–500.

Hedges, L.V. (2009). Statistical considerations. In H. Cooper, L.V. Hedges, & J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 37–47). New York: Russell Sage Foundation.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Hunter, J.E., & Schmidt, F.L. (1990a). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*, 334–349.

Hunter, J.E., & Schmidt, F.L. (1990b). *Methods of meta-analysis: Correcting error and bias in research findings.* Beverly Hills, CA: Sage.

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

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Hunter, J.E., Schmidt, F.L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594–612.

Le, H., & Schmidt, F.L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods*, *11*, 416–438.

Le, H., Schmidt, F.L., Harter, J., & Lauver, K. (in press). Revisiting conceptualizations of measurement error in research: Is organizational commitment really different from job satisfaction? *Organizational Behavior and Human Decision Processes*.

Le, H., Schmidt, F.L., & Putka, D.J. (2009). The multi-faceted nature of measurement error and its implications for measurement error corrections. *Organizational Research Methods*, *12*, 165–200.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, education, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *8*, 161–171.

MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Oh, I.-S. (2009). *The five factor model of personality and job performance in East Asia: A cross-cultural validity generalization study*. Doctoral dissertation, University of Iowa.

Oh, I.-S., Schmidt, F.L., Shaffer, J.A., & Le, H. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning & Education*, *17*, 563–570.

Ones, D.S. (2008, April). Generalizability of personality-expatriate performance relationships: Findings from the 10 GLOBE cultural clusters. In D.S. Ones & J. Deller (Chairs), *Expatriate success: Findings from ten host-cultural clusters around the world*. Symposium presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco.

Rothstein, H.R. (2003). Progress is our most important product: Contributions of validity generalization and meta-analysis to the development and communication of knowledge in I/O psychology. In K.R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 115–154). Mahwah, NJ: Erlbaum.

Salgado, J.F., Anderson, H., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, *56*, 573–605.

Schmidt, F.L. (1970). *The relative efficiency of regression and simple unit predictor weights in applied differential psychology*. Doctor dissertation, Purdue University, West Lafayette, IN.

Schmidt, F.L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, *31*, 699–714.

Schmidt, F.L. (1972). The reliability of differences between linear regression weights in applied differential psychology. *Educational and Psychological Measurement*, *32*, 879–886.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115–129.

Schmidt, F.L. (2008). Meta-analysis: A constantly evolving research integration tool. *Organizational Research Methods*, *11*, 96–113.

Schmidt, F.L., Berner, J.G., & Hunter, J.E. (1973). Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology*, *58*, 5–9.

Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540.

Schmidt, F.L., & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of statistical significance testing. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Hillsdale, NJ: Erlbaum.

Schmidt, F.L., & Hunter, J.E. (1999). Theory testing and measurement error. *Intelligence*, *27*, 183–198.

Schmidt, F.L., Hunter, J.E., & Outerbridge, A.N. (1986). The impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*, 432–439.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*, 473–485.

Schmidt, F.L., & Le, H. (2004). *Software for the Hunter–Schmidt meta-analysis methods* [Computer software]. Iowa City: Department of Management and Organization, University of Iowa.

Schmidt, F.L., Le, H., & Oh, I.-S. (2009). *The methodological and conceptual basis for estimation of correlations between constructs in classical measurement theory*. Manuscript submitted for publication.

Schmidt, F.L., Ocasio, B.P., Hillery, J.M., & Hunter, J.E. (1985). Further within setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology*, *38*, 509–524.

Schmidt, F.L., Oh, I.-S., & Hayes, T.L. (2009). Fixed vs. random models in meta-analysis: Model properties and comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128.

Schmidt, F.L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology*, *59*, 281–305.

Schmidt, F.L., Oh, I.-S., & Shaffer, J.A. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, *61*, 827–868.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3):25–32.

Thompson, B. (2007). Effect sizes and confidence intervals for effect sizes. *Psychology in the Schools*, *44*, 423–432.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.