Effect-Size Magnitude Benchmarks: Implications for Scientific Progress and Statistical Inferences

Authors

Frank A. Bosco, Virginia Commonwealth U., aom@frankbosco.com
Kulraj Singh, U. of Memphis, ksingh1@memphis.edu
James G Field, Virginia Commonwealth U., j.g.field@hotmail.com
Charles A. Pierce, U. of Memphis, capierce@memphis.edu

Effect-Size Magnitude Benchmarks:

Implications for Scientific Progress and Statistical Inferences

ABSTRACT

It has been a half-century since Cohen (1962) provided rough benchmarks for classifying effect size magnitude. However, such general benchmarks are not necessarily applicable to management research. By applying an innovative data collection protocol, we derive a set of benchmarks for the organizational behavior and human resource management (OB/HRM) research context by content-analyzing more than 100,000 effect sizes reported in *Journal of Applied Psychology* (*JAP*) or *Personnel Psychology* (*PPsych*) from 1980-2010. Results indicate that Cohen's (1988) benchmarks exhibit very little overlap with OB/HRM findings, with tertile partitions at $|r| = .09$ and .25. In addition, we classified the 100,410 effect sizes according to a hierarchical variable taxonomy to present several within-content domain, context-specific benchmarks for variable relations of different types. Broadly, results indicate that relations involving behaviors exhibit relatively small effect sizes with tertile partitions at roughly $|r| = .10$ and .25 and relations not involving behaviors exhibit larger effect sizes with partitions at roughly $|r| = .20$ and .40. We discuss implications for scientific progress, practical significance, Bayesian inference, and *a priori* power analysis.

Keywords:

Effect size; philosophy of science; meta-analysis

Effect-Size Magnitude Benchmarks:

Implications for Scientific Progress and Statistical Inferences

The necessity of effect size (ES) has been realized the APA Task Force on Statistical Inference, who noted that, for authors, provision of ES information is "almost always necessary" (APA, 2010: 34). Effect size plays an especially central role throughout the OB/HRM research process, including study design (e.g., *a priori* power analysis; hypothesis development) statistical analysis (e.g., meta-analysis; Bayesian techniques; Kruschke, Aguinis, & Joo, 2012), and estimates of practical significance (Aguinis et al., 2010) and scientific progress (Cohen, 1988; Ozer, 1985). Indeed, as Cohen (1988: 532) stated, "a moment's thought suggests that [ES] is, after all, what science is all about." Thus, it is clear that ES interpretation is central to understanding scientific progress and the value of scientific findings. However, the key issue is that currently used ES benchmarks (Cohen, 1988) were developed non-empirically in a different field-level research context (for a historical account, see Aguinis & Harden, 2009). Researchers thus face several challenges regarding ES consciousness or awareness that substantially limit ES usefulness (Vacha-Haase & Thompson, 2004).

The first challenge is that there are no existing omnibus ES benchmarks that provide a tailored frame of reference for the OB/HRM research context. That is, OB/HRM researchers have no empirical benchmarks to classify ESs in terms of their magnitude. As a common surrogate, Cohen's (1988) benchmarks define $|r|$ = .10, .30, and .50 as small, moderate, and large ESs, respectively. However, several researchers have noted that Cohen's (1988) benchmarks present with unrealistically high cutoff values (Hemphill, 2003). In addition, as Cohen (1988) himself cautioned, his benchmarks are non-empirical, based on a single volume of 1960s psychology research, and should be used only when no others are available. A second challenge regarding ES interpretation involves the existence of only one benchmark "lens." Instead, ES benchmarks should be tailored to particular research contexts (Aguinis & Harden, 2009; Hemphill, 2003). As an example, OB/HRM researchers and psychologists have long observed that ESs involving behaviors (e.g., job satisfaction-turnover) tend to be weaker than those involving intentions (e.g., job satisfaction-turnover intention) (Wicker, 1969). However, ES benchmarks for different relation types are not currently available in OB/HRM research, making it difficult to interpret ESs in context. Thus, what might be a relatively "large" ES in one context might be relatively "small" in another.

Several researchers have urged for a shift in the way we think about findings and that greater emphasis be placed on ESs and their confidence intervals (Cohen, 1988). Indeed, OB/HRM researchers have traditionally relied on null hypothesis significance testing (NHST) and reported findings in terms of whether statistical significance is achieved (Harlow, Mulaik, & Steiger, 1997). In addition, increased ES awareness and reporting have occurred alongside the increased application of meta-analytic techniques (e.g., Cumming, Fidler, Kalinowski, & Lai, 2011). Importantly, meta-analyses are more frequently cited among academics (Carlson & Ji, 2011), and more frequently reach practitioner audiences (Aguinis et al., 2011) than individual primary studies. A key opportunity to increase the practical relevance of OB/HRM research for practitioners is to consider alternative ways to describe and communicate our findings (e.g., ESs).

Researchers lament that the field of OB/HRM continues to exhibit a science-practice gap (e.g., Cascio & Aguinis, 2008) – that we *don't matter* to practitioners (Hambrick, 1994). Indicators of the gap's existence include decades of academics' suggestions on how best to narrow the divide (Aguinis et al., 2010) and low practitioner exposure to management research outlets (Rynes, Colbert, & Brown, 2002). The science-practice gap is especially troublesome because improving managerial practice through evidence-based techniques is central to the Academy of Management's mission statement (Aguinis et al., 2010). Shapiro, Kirkman, and Courtney's (2007) findings indicate that the *lost in translation* gap is more prevalent than the *lost before translation* gap. Thus, the field of management appears to suffer from a greater communication problem than a relevance problem, *per se*. We submit that this is relatively good news for practitioners and academics. Indeed, consumers can think about, summarize, and communicate existing research findings (e.g., ESs) in different ways without requiring a new corpus of research findings on a potentially different set of phenomena.

The purpose of the present manuscript is to present solutions to several key challenges associated with the definitions of ES magnitude in OB/HRM research. Specifically, we apply an innovative data collection protocol that allows us to extract a finer-grained set of 20 ES benchmarks for common relation types in OB/HRM research from more than 100,000 published ESs. We address existing benchmarks' generalizability to the OB/HRM context, ES interpretation, and sample sizes needed to achieve adequate power in a variety of common research contexts (i.e., relation "types").

Our manuscript is organized as follows. First, we discuss the role of ES benchmarks in OB/HRM research. We discuss limitations of Cohen's (1988) benchmarks and present examples of field-level and finer-grained benchmarks that have been tailored to other disciplines in the social sciences. Second, we report the results of a study in which we content-analyze 100,410 ESs published in *Journal of Applied Psychology* or *Personnel Psychology* from 1980-2010. Using a hierarchical taxonomy of OH/HRM variables, we categorize each effect size in terms of its variables' relation "type." As an example, at the taxonomy's broadest level of abstraction, job satisfaction-employee performance is considered an attitude-behavior relation. We provide ES and sample size benchmarks for very broad (e.g., all attitudes → all behavior), as well as more specific (e.g., attitudes toward people → performance) relations. In short, we provide an empirically-based set of 20 ES magnitude benchmarks that will serve as an alternative approach for ES interpretation, identifying areas of relative scientific success and failure, estimating practical significance, and applying statistical techniques (e.g., Bayesian inference; *a priori* power analysis).

GENERALIZABILITY OF EXISTING EFFECT-SIZE BENCHMARKS

With repeated ES consciousness-raising attempts (e.g., Cohen, 1988, 1992; Thompson, 2001, 2002) came benchmarks (i.e., guidelines; operational definitions) to interpret their relative magnitude. Cohen's (1988) effect size benchmarks (e.g., $|r|$ = .10, .30, and .50 as small, moderate, and large ESs, respectively) are widely used. The widespread prevalence of Cohen's (1988) benchmarks was noted by Hill et al. (2008: 177), in that "if there is any norm, it is to refer to Cohen's (1988) rules of thumb." The prevalence is likely due to the benchmarks' wide applicability – from interpreting ESs among others in a corpus of findings (e.g., Zakzanis, 2001), assessing scientific progress (Cohen, 1988), and specifying anticipated ESs during *a priori* power analysis for relatively nascent relations (e.g., Vacha-Haase & Thompson, 2004). Thus, as a starting point, one key question that OB/HRM researchers should ask is: Do existing ES benchmarks accurately represent the corpus of findings in OB/HRM research?

Although Cohen's (1988) benchmarks have raised ES consciousness, researchers in several fields have questioned their generalizability, as Cohen (1988) indeed cautioned. The concern has been raised in psychological assessment and treatment (Hemphill, 2003), neuropsychology (Zakzanis, 2001), educational psychology (Hill et al., 2008), software engineering (Kampenes et al., 2007), and others. Indeed, as Zakzanis (2001: 664) noted, "unless we begin to give due consideration to effect sizes in our neuropsychological studies, it will be difficult to determine how large one's obtained effects are within a specific domain of neuropsychological research without a frame of reference." Thus, an awareness of appropriate field-level ES benchmarks is essential.

As an example of field-level ES benchmarks, Kampenes et al.'s (2007) analysis of 92 software engineering experiments indicates a context-specific ES distribution split into thirds (i.e., small; moderate; large) at r_{pb} = .19 and .46. Although Kempenes et al. (2007) describe their benchmarks as similar to Cohen's (1988), the two rubrics would result in approximately 20-25% ES classification disagreement (see Kampenes et al., 2007, Table 8). As another example, Ellis (2010) observed a weighted mean r = .06 among 23 existing meta-analytic ES estimates in international business. Finally, Hemphill (2003) provided ES benchmarks for the areas of psychological assessment and treatment. From an analysis of two large meta-analytic datasets (Lipsey & Wilson, 1993, k = 302; Meyer et al., 2001, k = 78), Hemphill (2003) reported an ES distribution split into thirds at $|r|$ = .18 and .30, a substantial departure from Cohen's (1988) benchmarks split at $|r|$ =.30 and .50. Indeed, all effect sizes classified as moderately-sized by Cohen's (1988) benchmarks are large by Hemphill's (2003) standards. In short, reports in various scientific disciplines indicate the awareness of a large gap between traditional and tailored ES benchmarks and the benefits of their updating.

One reason for the lack of overlap between Cohen's (1988) and Hemphill's (2003) benchmarks lies in how the ES distributions are partitioned. Hemphill's (2003) benchmarks are presented as thirds (i.e., the 33rd and 67th percentiles represent lower bounds for moderate and large ESs, respectively). In contrast, Cohen's (1988) benchmarks are not to be interpreted as thirds, at least in the social sciences. Indeed, Cohen (1988: 78) noted that behavioral scientists including personnel psychologists rarely encounter substantive ESs greater than .50 or .60 – an "effective *upper limit*" (italics added). Thus, Cohen's (1988) benchmark for a large ES is best understood as an exceptionally large ES in the social sciences, and not simply one that would fall into the top third of the ES distribution. A similar problem is encountered with Cohen's (1988) definition of a *small* ES. Indeed, with a lower limit placed at $|r| = .10$ for a small ES, should we consider $|r| = .09$ as non-existent or zero? Thus, an additional drawback of Cohen's (1988) benchmarks is a lack of percentile definitions for the ES values, an issue that hampers the interpretation of ES and the development of an ES frame of reference (Zakzanis, 2001).

WHAT IS THE DISTRIBUTION OF EFFECT SIZES IN OB/HRM RESEARCH?

Two large datasets provide a starting point for estimating the field-level ES distribution for organizational (i.e., including OB/HRM) research. Aguinis, Dalton, Bosco, Pierce, & Dalton's (2011) ES distribution of more than 5,000 meta-analytically derived ESs in organizational research (e.g., *Academy of Management Journal*; *Journal of Management*; *Journal of Applied Psychology*) is split into thirds at $|r| = .18$ and .34, close to Hemphill's (2003) markers at $|r| = .18$ and .30. A similar analysis based on Dalton et al.'s (2012, Study 5) findings from over 2,000 published and non-published primary study ESs, in a similarly diverse sample of organizational research, results in benchmark splits at $|r| = .13$ and .28. Thus, especially noteworthy is the lack of overlap between Cohen's (1988) moderate ES range and those derived from large, recent reviews of the psychological and organizational research literatures. Indeed, it appears that what Cohen (1988) has defined as *moderate* is best described as *large* in the extant organizational literature. We submit that these observations serve as a wakeup call for the need to tailor a set of ES benchmarks to the OB/HRM research context.

HOW MANY ES BENCHMARKS?

Beyond specifying ES benchmarks at the field level, researchers have taken steps to delineate ESs benchmarks *within* fields. Put differently, some fields have at their disposal an understanding of ES magnitudes in varying contexts "relevant to the intervention, target population, and outcome being measured" (Hill et al., 2008: 172). The key issue is that the same ES (e.g., $|r| = .20$) could have a very different meaning in different contexts. Indeed, as noted by Hemphill (2003: 79), "Large and substantive reviews of the psychological research literature undoubtedly would reveal the importance of having different sets of…guidelines for different areas of investigation." As examples, we submit that staffing researchers would benefit from distinct benchmarks for predictive versus concurrent validity studies, or attitude-behavior versus attitude-intention relations. For researchers, each systematic relation between ES magnitude distribution and context (e.g., attitudes vs. behaviors as outcomes; Wicker, 1969) should also serve as an indication that finer-grained ES benchmarks are needed. Indeed, smaller ESs regarding behaviors compared to attitudes may actually be more practically important due to their relative proximity to outcomes that matter for individuals, organizations, and society.

As an example of within-field benchmarks from education research, Hill et al. (2008) observed that ESs (e.g., from classroom intervention studies) vary systematically as a function of student education level. Indeed, standardized test score gains between adjacent grade levels appear to decline substantially from the earliest comparison (i.e., K-1st grade; $d = 1.33$) to later comparisons (i.e., 11th-12th grade; $d = .04$) (Hill et al., 2008). As Hill et al. (2008: 174) note, "a particular [ES] from an intervention study, [$d = .10$], would constitute a relatively smaller substantive change for students in early grades than for students in later grades." Thus, the obtained ES must be interpreted within the particular grade level context. Hill et al. (2008)

provide similar context-dependent ES benchmarks for demographic performance gaps and various methodological choices (e.g., broad vs. narrow standardized tests as indicators of achievement). As Hill et al. (2008: 177) note, "when it comes to such findings, we thus conclude that one effect size rule of thumb does not and cannot fit all."

The social sciences abound with similar examples of the need for more specialized ES benchmarks. As noted by Wicker (1969: 65), ESs for attitude-behavior relations are “rarely above .30, and often are near zero... only rarely can as much as 10% of the variance in overt behavioral measures be accounted for by attitudinal data.” Similarly, Schwarz (2007: 638) stated that attitude-behavior ES magnitudes are “less than impressive.” Indeed, many attitude-behavior relations in OB/HRM research such as job satisfaction-job performance are small (e.g., r = .18, Judge, Thoresen, Bono, & Patton, 2001). However, intention-behavior relations such as turnover intention – turnover behavior can be much larger (e.g., r = .35; Griffeth, Hom, & Gaertner, 2000), and attitude-attitude relations such as overall job satisfaction–leader-member exchange (r = .46; Gerstner & Day, 1997) are often larger still (cf. Ajzen, 1987). As another example of the need for more refined benchmarks, consider a recent bare-bones meta-analytic estimate of the general mental ability - employee performance relation, r = .28 (cf. Schmidt, Shaffer, & Oh, 2008). Although generally considered one of the best predictors of employee performance (Ree & Earles, 1992), by Cohen’s benchmarks, the uncorrected ES represents a small effect size (i.e., $|r| < .30$). Thus, by extension, most predictors of employee performance (e.g., personality; general mental ability) are likely described as “small,” a rubric that does little to distinguish among obtained ESs, indicate scientific progress, and attract practitioner attention.

As an additional illustration, take two recent uncorrected meta-analytic estimates from the recruiting literature: compensation-job attraction (r = .14) and compensation-job choice (r =

.14) (Chapman, Uggerslev, Carroll, Piasentin, & Jones, 2005). From these findings, should we conclude that compensation is similarly related to job attraction (an attitude) and actual job choice (a behavior)? The answer is incomplete without a consideration of context and lack of context can lead to specious conclusions. Specifically, at r = .14, compensation is Chapman et al.'s (2005) *weakest* predictor of job attraction. In contrast, at r = .14, compensation ranks above the median of actual job choice predictors. In short, although omnibus ES benchmarks provide information on the distribution of *all possible* relations' ESs, availability of more specific benchmarks would allow for more accurate within-context comparisons.

PRESENT STUDY

The present study provides a large-scale analysis of OB/HRM research from a database of more than 100,000 ESs published in *Journal of Applied Psychology* or *Personnel Psychology* from 1980-2010. From analyses of the ESs coded according to a hierarchical variable taxonomy, we approach two central research questions. First, we ask: To what extent do Cohen's (1988) ES benchmarks generalize to the OB/HRM research context? To answer this question, we present the most comprehensive set of field-level, omnibus ES benchmarks for OB/HRM researchers and contrast them with existing benchmarks. As a second research question, we ask: Are common bivariate relation "types" in OB/HRM research associated with different ES distributions? To this end, we provide ES benchmarks for 20 common relation "types" in OB/HRM research (e.g., attitude-attitude vs. attitude-behavior relations), and describe how more refined benchmarks can better inform *a priori* power analysis. Taken together, we provide an empirically-based understanding of ES distributions in OB/HRM research –broadly and in particular contexts- that can be used to assess scientific progress, estimate practical significance, and inform a host of data-analytic techniques (e.g., *a priori* power analysis, Bayesean inference).

METHOD

Database

We collected all correlation coefficients reported in tables of *Journal of Applied Psychology* (*JAP*) and *Personnel Psychology* (*PPsych*) from 1980 to 2010 with the use of custom software. A total of approximately 150,000 ESs and their respective sample sizes are included in the database. For the present study, we randomly selected articles from the full database and coded each article's ESs for their variables' "types" until we had reached 100,000 ESs. Analyses in the present manuscript are conducted at the ES unit of analysis, and based on a total of 100,410 ESs.

To code for variable type, three of the authors created a hierarchical taxonomy of variables published in tables of *JAP* and *PPsych* articles from 1980-2010. To this end, the first author created rough taxonomy from existing typologies in OB/HRM research (Cascio & Aguinis, 2008; Crampton & Wagner, 1994). Next, we followed the approach by Aguinis, Pierce, Bosco, and Muslin (2009) and refined the taxonomy through several rounds of error checks and discussions with the third and fourth authors. As an example, attitudes are categorized in terms of their respective targets (e.g., attitudes toward the job; toward people; toward the organization). Similarly, behaviors are categorized in terms of their major "types" (e.g., performance; turnover). The taxonomy is comprehensive, and covers all major topics found in OB/HR textbooks. In total, the taxonomy arranges 4,553 nodes (i.e., variables or variable branches) into 10 first-level nodes (e.g., behaviors; attitudes; intentions), which then branch to a mean of 5.2 second-level nodes (e.g., behaviors: performance; behaviors: turnover), and so forth.

In contrast to existing typologies (e.g., Cascio & Aguinis, 2008) the present taxonomy of variables presents a taxonomic display pertaining to *what variables are* rather than *how they are* *used*. As an example, although personality traits are categorized as a predictor of employee performance in existing typologies (Cascio & Aguinis, 2008), they are also used as a predictor of employee turnover and many other organizationally-relevant outcomes (Zimmerman, 2008). Thus, as an example, in the present taxonomy, personality traits are categorized more broadly under the first-level node: person characteristics. An abbreviated version of the taxonomy contents is displayed in Figure 1.

The present analyses are based on a database of 100,410 ESs from 1,277 unique articles containing 19,803 variable entries. Thus, articles contain a mean of 78.63 ESs, or roughly the equivalent of one 13-variable correlation matrix. Note that many articles contain more than one correlation matrix (i.e., to present findings for multiple samples or studies). However, because the present analyses are conducted at the ES unit of analysis (i.e., to make inferences regarding a corpus of ESs), sample dependence does not threaten the validity of our inferences as it might in a traditional, substantive meta-analysis (see Aguinis et al., 2011 for a similar rationale).

Coding Process and Agreement

The third and fourth authors coded all variables in the dataset according to the taxonomy. Thus, for each of the 19,803 rows of data, only one piece of information was coded: a unique identifier (i.e., five-digit code) from the variable taxonomy corresponding to the particular variable node. As an example of the hierarchical classification, the variable leader-member exchange (LMX; Gerstner & Day, 1997) is located in the taxonomy as a fifth-level node (i.e., all attitudes → attitudes toward people → attitudes toward supervisors/mentors → relationship quality → LMX). Coders used a combination of exact letter sting matching with the taxonomy's node text and decision-making to code each variable. Variables with very low frequency of occurrence (e.g., prejudicial attitudes against West Germans, Petersen & Dietz, 2008) were coded as miscellaneous by assigning a broad classification node (e.g., attitudes toward people).

After the 19,803 variables were coded according to the taxonomy, database tools in Microsoft Excel were used to create the list of 100,410 ESs with taxonomy node codes for each variable in the pair. As an example, if a given correlation matrix contained 13 variables, only the 13 variables' taxonomic assignments required manual coding. From these 13 codes, a total of 78 bivariate relations "code pairs" were produced and linked to the ES and sample size information (see Bosco, Field, & Pierce, 2012) in the database using range lookup formulas.

To assess coder agreement, articles were randomly selected until each coder had independently coded at least 300 ESs (301 ESs were coded). For the present study, we assess agreement at broad levels of categorization. As an example, although LMX is coded as a fifth-level node, the present agreement assessment is based on third-level or broader classifications (e.g., LMX = attitudes toward people). We assessed agreement at this broad level of abstraction because the goal of the present manuscript is to provide ES distributions for broad classifications of variables. Thus, our agreement assessment method is aligned with our present analyses, inferences, and proposed applications. The two coders agreed on 278 (92.4%) of the 301 assignments, but due to database's size, disagreements were not resolved.

RESULTS

Omnibus Field-level Benchmarks

Our first research question asks to what extent existing ES benchmarks accurately reflect the extant OB/HRM literature. To this end, we provide parameters that describe the distribution of 100,410 ESs in our database. We summarize the distribution with two primary analytic approaches. First, we provide percentiles to partition the distribution into between two and five equal parts (i.e., 20th, 25th, 33rd, 40th, 50th, 60th, 67th, 75th, and 80th percentiles). Second, we provide a bare-bones meta-analytic estimate for the 100,410 ESs.

As shown in Table 2, the distribution of 100,410 ESs exhibits a "small" (Cohen, 1988) median ES, $|r|$ = .16, and is split into thirds (i.e., upper and lower boundaries for moderate ES) at $|r|$ = .09 and .25. Our observed moderate ES range is thus similar to Dalton et al.'s (2012) ES distributions split into thirds at $|r|$ = .10 and .22 (published ESs) and $|r|$ = .11 and .28 (non-published ESs), but substantially different (i.e., non-overlapping moderate ES range) when compared to Cohen's (1988) benchmarks at $|r|$ = .30 and .50 (see Figure 2). In addition, as shown in Table 2, we observed values of $|r|$ = .05, .07, .09, .12, .16, .21, .25, .31, and .36 for the 20th, 25th, 33rd, 40th, 50th, 60th, 67th, 75th, and 80th percentiles, respectively. Cohen's (1988) benchmarks for small, moderate, and large ESs (i.e., $|r|$ = .10, .30, .50) correspond to the 35th, 73rd and 90th percentiles, respectively, of our distribution of 100,410 ESs.

As a second analytic approach to summarizing the distribution of the 100,410 ESs, we conducted a bare-bones meta-analysis. As shown in Table 3, our analysis revealed a "small" mean ES ($|r|$ = .220; 95% CI = .218, .221; k = 100,410; N = 197,171,724). The unweighted mean ES revealed a similar value, $|r|$ = .218. As might be expected with a large, diverse collection of ESs, our results indicate that moderation is likely. Indeed, as shown in Table 3, the I^2 statistic (Higgins & Thompson, 2002) approaches its maximum value of 100 in the present dataset (I^2 = 98.95), and the 80% credibility interval (-.05, .22) includes zero (cf. Hunter & Schmidt, 2004).

We note that the median ES value reported above, $|r|$ = .16, is smaller than the mean meta-analytically derived ES, $|r|$ = .22, indicating that the distribution of ESs is positively skewed (skew = 1.28). A positively-skewed ES distribution was expected for two reasons. First, as noted by Cohen (1988), "large" ESs are relatively rare in social science research. Second, as in large surveys of ESs in organizational research (e.g., Dalton et al., 2012), we analyzed absolute ES rather than raw ES. This distinction is noteworthy because an analysis of ESs from any particular relation (i.e., those including any negative ESs) should reveal a more approximately normal distribution. Indeed, untransformed ES values for one of OB/HRM's largest meta-analyses on a single topic (Judge et al., 2001, job satisfaction–job performance; k = 312) reveals skew = 0.73 for r and skew = 1.31 for $|r|$; the latter value is almost identical to that obtained in the present analysis.

Context-specific ES benchmarks

Our second research question asks whether different major relation "types" exhibit distinct ES distributions. As described earlier, distinct *within-field* benchmarks for relations of different types or in different research contexts have been suggested (Hemphill, 2003) and provided (Hill et al., 2008) in the social sciences. To address our second question, we identified the most frequent, substantive bivariate relation types (e.g., psychological characteristics → performance) in our database using count functions in Microsoft Excel, with the limitation that at least 200 ESs are associated with the bivariate relation type. We identified 20 common, broad bivariate relation types. Note that several categorizations contain overlapping ES sets (e.g., *performance* and *turnover* are subsets of *behavior* in our taxonomy). We follow the same analytic approach used to answer our first research question. Specifically, we present ES values at percentiles needed to split each group of ESs into between two and five equal groups (see Table 2). In addition, we provide bare-bones meta-analytic values for each bivariate relation type (see Table 3). Finally, Table 4 presents sample sizes needed to achieve .80 power *a priori* (cf. Cohen, 1988) for each relation type. Although we present values to achieve .80 power only, we present the inputs needed to estimate any level of power.

The variable types included in the 20 bivariate relation types (i.e., one omnibus ES distribution, and 19 more specific distributions) are shown with examples in Table 1. The 19 more specific distributions for which at least 200 ESs were available are arranged as follows, in three levels of abstraction (i.e., coarse, fine, and extra fine relation types). We identified four coarse relation types: attitudes-attitudes, attitudes-intentions, attitudes-behaviors, and intentions-behaviors. In addition, we identified four fine bivariate relation types with performance behavior (attitudes-performance; KSAs-performance; psychological characteristics-performance; objective person characteristics-performance), and three extra fine relation types for the attitudes-performance relation type (organization attitudes-performance; job attitudes-performance; people attitudes-performance). Similarly, we identified three fine relation types with employee turnover behavior (attitudes-turnover; psychological characteristics-turnover; objective person characteristics-turnover), and two extra fine relation types for the attitudes-turnover relation (organization attitudes-turnover; job attitudes-turnover). Finally, we observed three fine relation types among attitude types (organization attitudes-job attitudes; organization attitudes-people attitudes; job attitudes-people attitudes).

As shown in Table 2, we observed substantial variance in ES distribution parameters across the 20 bivariate relation types. Specifically, the four coarse relation types provide definitions of "moderately sized" ESs with partitions at $|r| = .17$ and .38 (attitudes-attitudes), $|r| = .18$ and .38 (attitudes-intentions), $|r| = .10$ and .24 (attitudes-behaviors), and $|r| = .11$ and .26 (intentions-behaviors). Thus, for relations involving behaviors, ES values greater than roughly $|r| = .25$ exist in the upper third of the ES distribution (i.e., a "large" ES). In contrast, for coarse relations not involving behaviors (i.e., attitudes-attitudes; attitudes-intentions), the corresponding value for a "large" ES is roughly $|r| = .40$. Importantly, the distinction between broad relation types involving behaviors compared to those not involving behaviors is substantial. Indeed, our findings indicate that achieving 6.25% variance explained ($|r|$ = .25) when predicting behavior represents a "large" ES, but one needs to achieve 14.4% (i.e., $|r|$ = .38) for a "large" ES among non-behavioral relations (i.e., attitudes-attitudes; attitudes-intentions). These findings, especially for relations involving behaviors, indicate that Cohen's (1988) benchmarks do not accurately represent the OB/HR literature, and that tailored ES benchmarks are necessary.

Values for the three fine relation types with employee performance are shown in Table 2. Our findings reveal boundaries for a "moderate" ES at $|r|$ = .14 and .31 (KSAs-performance), $|r|$ = .10 and .23 (psychological characteristics-performance), $|r|$ = .05 and .13 (objective person characteristics-performance), and $|r|$ = .10 and .25 (attitudes-performance). In addition, three extra fine relations within the attitudes-performance relation type reveal moderate ES partitions at $|r|$ = .10 and .21 (organization attitudes-performance), $|r|$ = .09 and .25 (job attitudes-performance), and $|r|$ = .15 and .36 (people attitudes-performance). Thus, our findings reveal that KSAs are more strongly related with performance than attitudes (broadly) and psychological characteristics. In addition, objective person characteristics exhibit relatively weak relations with performance.

The three fine relation types with turnover behaviors are presented in Table 2. Our findings indicate moderate ES partitions at $|r|$ = .06 and .16 (psychological characteristics-turnover), $|r|$ = .05 and .11 (objective person characteristics-turnover), and $|r|$ = .08 and .20 (attitudes-turnover). In addition, two extra fine relation types for the attitudes-turnover relation type revealed moderate ES partitions at $|r|$ = .07 and .17 (organization attitudes-turnover) and $|r|$ = .11 and .22 (job attitudes-turnover). Thus, our findings reveal that turnover behavior is predicted relatively poorly compared to performance behavior. In addition, broadly, relations with employee turnover behavior larger than $|r|$ = .20 exist within the top third of the ES distribution (i.e., “large” ES).

In sum, results from over 100,000 ESs published from 1980-2010 in two leading OB/HRM journals indicate that commonly used, existing ES benchmarks (e.g., Cohen, 1988) are not appropriately tailored to the OB/HRM research context. In addition, results indicate that operational definitions for ES magnitude vary as a function of the bivariate relation type. Thus, the present study provides the first and largest field-level (and finer-grained) analysis of primary study ESs with distribution characteristics. Our findings indicate that interpretation, communication, and methodological choices involving ES require revision.

DISCUSSION

As Hill et al. (2008: 177) noted, in contrast to relatively clear-cut interpretation rules regarding the statistical significance of findings, the interpretation of ES “does not benefit from such theory or norms.” Indeed, as described earlier, many researchers in the social sciences have relied on a single ES benchmark lens for interpretation – Cohen’s (1988) benchmarks. As shown in Figure 2, results of the present study indicate that the ES benchmark generalizability concern originally raised by Cohen (1988) and echoed by others (e.g., Hemphill, 2003; Hill et al., 2008) is well-founded. In addition, the present results indicate that “one [ES] rule of thumb does not and cannot fit all” (Hill et al., 2008: 177) in the OB/HRM research context.

Our first research question addresses the extent to which Cohen’s ES benchmarks reflect the distribution of findings in OB/HRM research. Our analyses of more than 100,000 ESs indicate that Cohen’s (1988) ES benchmarks exhibit very little overlap with the present observed distribution. Indeed, whereas Cohen (1988) defined minimum values of moderate and large ESs as $|r|$ = .30 and .50 (i.e., 9 and 25% variance explained, respectively), our findings indicate that moderate and large ESs in OB/HRM research, operationalized as tertiles (e.g., Hemphill, 2003), are rather on the order of $|r| = .09$ and .25 (i.e., 3 and 6.25% variance explained, respectively). Thus, in terms of variance explained, to interpret ESs in OB/HRM research according to Cohen's (1988) benchmarks is erroneous by a factor of three. Distinctions like these have potentially enormous detrimental effects on ES consciousness and the science-practice gap.

The benefits of a recalibrated set of omnibus, field-level ES benchmarks are relatively clear. Indeed, as other researchers have described, "many of these other guidelines by which the magnitude of correlation coefficients are compared are unrealistically large and "inappropriate" (Meyer et al., 2001: 132)" (Hemphill, 2003: 78). However, crucially, the lack of calibration leads to erroneous thinking and communication of research findings. As an example, take a recent uncorrected meta-analytic estimate of the general mental ability (GMA)-performance relation ($r = .28$, cf. Schmidt et al., 2008). In addition, consider the most recent uncorrected meta-analytic estimates for the Big Five personality traits: conscientiousness ($r = .14$), emotional stability ($r = .09$), agreeableness ($r = .07$), extraversion ($r = .06$), and openness to experience ($r = .04$) (Hurtz & Donovan, 2000). Although it is clear that personality traits explain relatively little variance in performance compared to GMA, consumers of science with Cohen's (1988) benchmark lens are forced to conclude that all of these predictors exhibit "small" or non-existent effects (i.e., $|r| < .30$, $|r| < .10$; Cohen, 1988).

In contrast, by interpreting the meta-analytic findings above according to the present recalibrated ES distribution (i.e., omnibus tertile partitions at $|r| = .09$ and .25), we submit that three of the Big Five personality traits (agreeableness, extraversion, and openness to experience) present with "small" ESs (i.e., lower tertile; $|r| < .09$) when compared to a sample of "all" possible other relations in the OB/HRM context. In addition, at $r = .09$, emotional stability meets the minimum qualification for classification as a moderate ES, with conscientiousness performing slightly better (r = .14). However, according to the present benchmarks, GMA presents with a "large" ES (r = .28; Schmidt et al., 2008), a classification compatible with the statement that "intelligence is the best predictor of job performance" (Ree & Earles, 1992: 86). Indeed, to say that GMA is one of the best predictors of job performance and also a "small" ES approaches meeting the criteria for classification as a non sequitur, and does little to bridge the science-practice gap (cf. Cascio & Aguinis, 2008).

Our second research question addresses the extent to which benchmarks vary across *types* of bivariate relations (e.g., attitude-attitude vs. attitude-behavior). Results indicate that substantial variance in the tertile partitions exists across relation types and, thus, one single benchmark will not do (see Table 2). For heuristic purposes, relations involving behaviors (i.e., attitudes-behaviors; intentions-behaviors) exhibit tertile splits at roughly $|r|$ = .10 and .25. In addition, for broad relation types not involving behaviors (e.g., attitudes-attitudes; attitudes-intentions), tertile splits are placed at roughly $|r|$ = .20 and .40. As shown in Table 2, the distribution parameters also vary to some extent at the fine and extra-fine level of abstraction. As examples, attitudes toward people tend to explain more variance in performance than attitudes toward either the organization or job. Similarly, attitudes toward the job tend to predict turnover behavior to a slightly greater degree than attitudes toward the organization.

Implications for Statistical Approaches

Our results have important implications for several research process stages. During the research design stage, an anticipated ES is specified to conduct *a priori* power analysis (a value that informs the data collection phase). While existing meta-analytic estimates and/or direct replications represent suitable sources, Cohen's (1988) benchmark approach (e.g., anticipating a small, moderate, or large ES; |*r*| = .10, .30, or .50) should be a last resort approach (Cohen, 1988). As a route to partially ameliorating this "most difficult part of power analysis" (Cohen, 1992: 156), we offer an alternative approach for nascent research areas: anticipated ES specification based on broad relation *types*. Indeed, *a priori* power analysis should rely on the most context-specific ES benchmarks available (Hill et al., 2008). However, when a precise value is not available, researchers would be better served to specify a *typical* context-specific ES (e.g., for an attitude-behavior relation) rather than to make a "shot in the dark" with Cohen's (1988) benchmarks. As described earlier, our findings indicate that Cohen's (1988) benchmarks present unrealistically high values for the OB/HRM research context, the use of which could lead to upwardly biased ES forecasts and, thus, underpowered studies (cf. Maxwell, 2004).

As Cohen (1988; 1992) noted, *a priori* power analysis is an essential for research planning, and aids in the reduction of Type II errors. At the field level, the median sample size for ESs published in *JAP* from 1995-2008 is 173 (Shen et al., 2011). At a sample size of 173, only anticipated ESs greater than |*r*| = .21 would have achieved statistical power greater than .80 during *a priori* power analysis. The present analyses reveal that |*r*| = .21 is an ES that corresponds with the 60th percentile of the full ES database distribution. Indeed, our observed median (i.e., 50th percentile) ES, |*r*| = .16, would require 304 observations to achieve *a priori* power = .80. Using the two median values just described (i.e., N = 173; Shen et al., 2011) and |*r*| = .16 from the present analyses, the median finding in OB/HRM research is associated with *a priori* power = .56. Thus, it appears that OB/HRM research continues to suffer from lack of statistical power (Maxwell, 2004), and we submit that refined benchmarks will reduce this problem by providing a more realistic estimate of sample size needed to achieve statistical power in context.

As an additional implication for scientists, the present results inform hypothesis formation and data analytic techniques. Indeed, Edwards and Berry (2010) have argued that point estimates are preferable to null hypotheses, and Bayesian statistical techniques make the call for similar information: a prior distribution (Kruschke et al., 2012). Although Bayesian statistical approaches are rarely applied in OB/HRM research, their application in other scientific disciplines has been described as revolutionary and able to foster the development of a cumulative scientific program (Kruschke et al., 2012; Schmidt, 2008). However, specifying the prior distribution can be challenging. Indeed, as noted by Kruschke et al. (2012: 728), “the prior distribution is not capricious and must be explicitly reasonable to a skeptical scientific audience.”

Finally, our results have implications for the communication of findings and the estimation of practical significance. Indeed, researchers need to know where their findings stand relative to other, existing findings (i.e., ESs) in order to communicate them. With increased consciousness of ES distributions, researchers will be able to ask higher-order questions like, “overall, how well can we predict performance versus turnover?” Put differently, how well is our science able to predict key practitioner- and organizational performance-relevant outcomes? However, crucially, before such “big science” questions can be addressed, a field-level understanding of ES distributions is essential. In addition, a “big science” understanding of ES in context will allow for analyses at the meta-theoretical level. For example, researchers will be able to address questions like, “Do broad model classes (e.g., attitude→intention→behavior) show similar validity patterns across different OB/HRM-related behaviors?”

As another implication for scientific progress, increased consciousness of ESs observed across different major criteria would allow for the identification of research areas that tend to lag behind others in terms of explained variance. As an example, the present analyses indicate that, overall, researchers tend to have more success predicting employee performance compared to employee turnover. We submit that in order for scientific progress to be realized, we must first realize where progress is not being made. While the solution to an overall lack of scientific progress is beyond the scope of the present manuscript, we propose that field-level analyses can shed light on areas in need of more qualitative research (cf. Colquitt & Zapata-Phelan, 2007).

Limitations and Future Directions

First, we note that we have only summarized ESs found in tables of two prestigious OB/HRM journals (*JAP* and *PPsych*) from 1980-2010. It remains possible that ESs found in other journals from other points in time might reveal different distribution parameters. We note, however, that our analyses were based on the largest ES database in the field of OB/HRM and, although including a wider range of journal sources, other similar content analyses have involved substantially smaller datasets (e.g., Aguinis et al., 2011; Dalton et al., 2012; Shen et al., 2012). In addition, it remains possible that, by dint of including only "top-tier" journals, our ES estimates are upwardly biased (e.g., due to publication bias).

As another limitation, it is worth noting that our analyses are very coarse, and include thousands of ESs that were not hypothesized per se, but included in correlation matrices nonetheless. As an example, one need not look far to discover correlations within matrices representing the relation between the person variables age and sex. As such, the present ES distribution parameters could be downwardly biased. However, as described earlier, existing studies use a similar approach (Dalton et al., 2012) and the compromise position we present necessitates some degree of taxonomic coarseness.

As additional future research directions, researchers should consider the development of multiple ES benchmarks. For example, benefits would come from an understanding not only of distribution parameters for psychological characteristics as predictors of job performance, but also distributions associated with predictive versus concurrent validity studies in this context. As noted by Hill et al. (2008: 177), “indeed, it is often useful to use multiple benchmarks when assessing the observed impacts of an intervention.” We submit that such top-down categorizations of ES types and contexts would serve as a beneficial starting point to understand the cumulative nature of scientific progress, as well as inform Bayesian statistical approaches through the development of cumulative prior distribution data. Indeed, by amassing data relevant to multiple ES benchmarks, researchers could also gain an understanding of the relative influence of construct-based versus methods-based research decisions. Put differently, questions could include: What matters more - what is measured or how it is measured?

CONCLUSION

In the present study, we analyzed more than 100,000 ESs reported in *JAP* and *PPsych* published from 1980-2010. We develop a hierarchical taxonomy of variables that appear in the two journals during this timeframe, and conduct high-order analyses of ES distributions across 20 major bivariate relation types. Results indicate that existing ES benchmarks (e.g., Cohen, 1988) do not represent the OB/HRM research findings. Specifically, results indicate that the distribution of ESs observed exhibits tertile partitions at values approximately one-half to one-third those intuited by Cohen (1988). In addition, results indicate that a single, omnibus set of ES benchmarks, while informative, serves to restrict the potential usefulness of ES benchmarks in general. Our results have important implications for each of the stages of the research process, from study design (e.g., power analysis; more specific hypotheses) to the interpretation of results.

REFERENCES

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. 2011. Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. ***Journal of Management***, 37: 5-38.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. 2010. Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. ***Organizational Research Methods***, 13: 515-539.

Ajzen, I. 1987. Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. ***Advances in Experimental Social Psychology***, 20: 1-63.

American Psychological Association. 2010. ***Publication manual of the American Psychological Association (6th edition)***. Washington, DC: Author.

Beyth-Marom, R., Fidler, F., & Cumming, G. 2008. Statistical cognition: Towards evidence-based practice in statistics and statistics education. ***Statistics Education Research Journal***, 7: 20-39.

Bosco, F. A., Field, J. G., & Pierce, C. A. 2012. Accommodational plasticity in organizational science: Influence of hypothesis framing on effect size. ***Academy of Management Best Paper Proceedings***, 2012: 1-6.

Carlson, K. D., & Ji, F. X. 2011. Citing and building on meta-analytic findings: A review and recommendations. ***Organizational Research Methods***, 14: 696-717.

Cascio, W. F., & Aguinis, H. 2008. Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. ***Journal of Applied Psychology***, 93: 1062-1081.

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. 2005. Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. ***Journal of Applied Psychology***, 90: 928-944.

Colquitt, J. A., & Zapata-Phelan, C. P. 2007. Trends in theory building and theory testing: A five-decade study of the *Academy of Management Journal*. ***Academy of Management Journal***, 50: 1281-1303.

Crampton, S. M., & Wagner, J. A. III. 1994. Percept-percept inflation in microorganizational research: An investigation of prevalence and effect. ***Journal of Applied Psychology***, 79: 67-76.

Cohen, J. 1992. A power primer. ***Psychological Bulletin***, 112: 155-159.

Cohen, J. 1988. ***Statistical power analysis for the behavioral sciences*** (2nd edition). New Jersey: Lawrence Erlbaum Associates.

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

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. 2011. The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. ***Australian Journal of Psychology***, 64: 138-146.

Dalton, D. R., Aguinis, H., Dalton, C. A., Bosco, F. A., & Pierce, C. A. 2012. Revisiting the file drawer problem in meta-analysis: An empirical assessment of published and non-published correlation matrices. ***Personnel Psychology***, 65: 221-249.

Dybå, T., Kampenes, V. B., & Sjøberg, D. I. K. 2006. A systematic review of statistical power in software engineering experiments. ***Information and Software Technology***, 48: 745-755.

Ellis, P. D. 2010. Effect sizes and the interpretation of research results in international business. ***Journal of International Business Studies***, 41: 1581-1588.

Fritz, C. O., Morris, P. E., & Richler, J. J. 2012. Effect size estimates: Current use, calculations, and interpretation. ***Journal of Experimental Psychology: General***, 141: 2-18.

Gerstner, C. R., & Day, D. V. 1997. Meta-analytic review of leader-member exchange theory: Correlates and construct issues. ***Journal of Applied Psychology,*** 82: 827-844.

Griffeth, R. W., Hom, P. W., & Gaertner, S. 2000. A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. ***Journal of Management,*** 26: 463-488.

Hambrick, D. C. 1994. 1993 Presidential Address: What if the Academy actually mattered? ***Academy of Management Review***, 19: 11-16.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). 1997. ***What if there were no significance tests*****?** Mahwah, NJ: Erlbaum.

Hemphill, J. F. 2003. Interpreting the magnitudes of correlation coefficients. ***American Psychologist***, 58: 78-79.

Higgins, J. P. T., &Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. ***Statistics in Medicine***, 21: 1539-1558.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. 2008. Empirical Benchmarks for Interpreting Effect Sizes in Research. ***Child Development Perspectives***, 2: 172-177.

Hunter, J. E., & Schmidt, F. L. 2004. ***Methods of meta-analysis: Correcting error and bias in research findings (2nd edition)***. New York, NY: Academic Press.

Hurtz, G. M., & Donovan, J. J. 2000. Personality and job performance: The Big Five revisited. ***Journal of Applied Psychology***, 85: 869-879.

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. 2001. The job satisfaction-job performance relationship: A qualitative and quantitative review. ***Psychological Bulletin***, 127: 376-407.

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. K. 2007. A systematic review of effect size in software engineering experiments. ***Information and Software Technology***, 49: 1073-1086.

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. ***Personality and Social Psychology Review***, 2: 196–217.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. ***Organizational Research Methods***, 15: 722-752.

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

Maxwell, S. E. 2004. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. ***Psychological Methods***, 9: 147-163.

McGrath, R. E., & Meyer, G. J. 2006. When effect sizes disagree: The case of r and d. ***Psychological Methods***, 11: 386-401.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. 2001. Psychological testing and psychological assessment: A review of evidence and issues. ***American Psychologist***, 56: 128-165.

Oakes, M. 1982. Intuiting strength of association from a correlation coefficient. ***British Journal of Psychology***, 73: 51-56.

Ozer, D. J. 1985. Correlation and the coefficient of determination. ***Psychological Bulletin***, 97: 307-315.

Petersen, L., & Dietz, J. 2008. Employment discrimination: Authority figures' demographic preferences and followers' affective organizational commitment. ***Journal of Applied Psychology***, 93: 1287-1300.

Ree, M. J., & Earles, J. A. 1992. Intelligence is the best predictor of job performance. ***Current Directions in Psychological Science***, 1: 86-89.

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

Schmidt, F. L., Shaffer, J. A., & Oh, I. 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.

Schwarz, N. 2007. Attitude construction: Evaluation in context. ***Social Cognition***, 25: 638-656.

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

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. 2011. Samples in Applied Psychology: Over a Decade of Research in Review. ***Journal of Applied Psychology***, 96: 1055-1064.

Sun, S., Pan, W., & Wang, L. L. 2010. A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. ***Journal of Educational Psychology***, 102: 989-1004.

Thompson, B. 2001. Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. ***Journal of Experimental Education***, 70: 80-93.

Thompson, B. 2002. ‘‘Statistical,’’ ‘‘practical,’’ and ‘‘clinical:’’ How many kinds of significance do counselors need to consider? ***Journal of Counseling & Development***, 80: 64-71.

Thompson, B. 2002. What future quantitative social science research could look like: confidence intervals for effect sizes. ***Educational Researcher***, 31: 25-32.

Vacha-Haase, T., & Thompson, B. 2004. How to estimate and interpret various effect sizes. ***Journal of Counseling Psychology***, 51: 473-481.

Wicker, A. W. 1969. Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. ***Journal of Social Issues***, 25: 41-78.

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

Zakzanis, K. K. 2001. Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. ***Archives of Clinical Neuropsychology***, 16: 653-667.

Zimmerman, R. D. 2008. Understanding the impact of personality traits on individuals’ turnover decisions: A meta-analytic path model. ***Personnel Psychology***, 61: 309-348.

Table 1

Examples of Major Variable Types Included in the Variable Taxonomy

Variable Type	**Examples**
Organization attitudes	Organizational commitment; perceived organizational support; procedural justice
People attitudes	Supervisor satisfaction; coworker satisfaction; leader-member exchange
Job attitudes	Job satisfaction; autonomy perceptions; pay satisfaction
Intentions	Turnover intention; intent to accept a job offer; intent to participate in development
Behaviors	Performance; absenteeism; turnover
Performance	In-role performance; extra-role performance; training performance
KSAs	Job knowledge; decision-making skills; general mental ability
Psychological characteristics	Conscientiousness; psychological capital; core self-evaluation
Objective person characteristics	Age; gender; ethnicity; tenure; physical health indicators
Turnover	Voluntary turnover; involuntary turnover; work withdrawal

Table 2
Effect Size Distribution Percentiles for Broad Relation Types[a]

			Effect Size Distribution Percentile								
Relation Type	***k*[b]**	**N**	**20th**	**25th**	**33rd**	**40th**	**50th**	**60th**	**67th**	**75th**	**80th**
(All effect sizes)	100,410	197,171,724	.05	.07	.09	.12	.16	.21	.25	.31	.36
Attitudes : attitudes	10,068	4,527,323	.10	.13	.17	.21	.27	.34	.38	.45	.50
Organization attitudes : Job attitudes	1,145	519,329	.12	.15	.19	.23	.29	.35	.38	.43	.48
Organization attitudes : People attitudes	471	252,292	.14	.17	.23	.28	.34	.40	.43	.48	.52
Job attitudes : People attitudes	447	242,562	.08	.10	.14	.17	.23	.28	.30	.35	.37
Attitudes : intentions	1,222	556,225	.11	.14	.18	.22	.27	.33	.38	.44	.47
Attitudes : behaviors	6,348	2,998,404	.06	.07	.10	.12	.16	.20	.24	.29	.34
Intentions : behaviors	400	167,887	.06	.09	.11	.13	.18	.23	.26	.32	.35
Performance : attitudes	2,207	712,402	.06	.08	.10	.13	.16	.21	.25	.31	.35
Performance : organization attitudes	569	158,077	.07	.08	.10	.13	.16	.19	.21	.27	.30
Performance : job attitudes	949	367,870	.06	.07	.09	.11	.16	.20	.25	.32	.36
Performance : people attitudes	302	67,500	.09	.12	.15	.18	.24	.30	.36	.41	.46

Performance : knowledge, skills, & abilities	1,140	1,182,197	.08	.10	.14	.16	.22	.26	.31	.38	.44
Performance : psychological characteristics	2,680	735,327	.06	.08	.10	.12	.16	.20	.23	.28	.32
Performance : objective person characteristics	1,259	863,852	.03	.04	.05	.06	.08	.11	.13	.16	.19
Turnover : attitudes	466	376,534	.05	.06	.08	.10	.13	.17	.20	.24	.26
Turnover : organization attitudes	219	89,590	.04	.06	.07	.09	.11	.14	.17	.23	.24
Turnover : job attitudes	201	262,308	.06	.07	.11	.13	.16	.2	.22	.26	.28
Turnover : psychological characteristics	254	109,364	.04	.05	.06	.08	.10	.13	.16	.19	.20
Turnover : objective person characteristics	323	3,186,252	.02	.03	.05	.06	.07	.09	.11	.14	.16

[a]*k*: number of samples; N: number of observations.
[b]Samples are dependent.

Table 3
Bare-bones Meta-analytic Estimates for Broad Relation Types[a]

Relation Type	*k*	N	unwt mean *r*	wt mean *r*	SD*r*	95% CONF lower	95% CONF upper	80 % CRED lower	80% CRED upper	I^2
(All effect sizes)	100,410	197,171,724	.218	.220	.209	.218	.221	-.047	.221	98.95
Attitudes : attitudes	10,068	4,527,323	.306	.300	.203	.296	.304	.045	.555	95.54
Organization attitudes : Job attitudes[c]	1,140	412,265	.304	.299	.181	.289	.310	.075	.523	92.99
Organization attitudes : People attitudes	471	252,292	.338	.327	.198	.309	.344	.078	.575	96.18
Job attitudes : People attitudes	447	242,562	.245	.220	.157	.205	.234	.026	.414	93.23
Attitudes : intentions	1,222	556,225	.296	.279	.189	.269	.290	.044	.515	94.75
Attitudes : behaviors	6,348	2,998,404	.207	.176	.190	.172	.181	-.061	.414	94.51
Intentions : behaviors	400	167,887	.219	.201	.165	.185	.217	-.003	.404	91.99
Performance : all attitudes	2,207	712,402	.220	.163	.162	.156	.170	-.003	.359	88.81
Performance : organization-targeted attitudes	569	158,077	.196	.180	.155	.167	.193	-.004	.364	85.96
Performance : job-targeted attitudes	949	367,870	.222	.141	.148	.131	.150	-.037	.319	88.60
Performance : people-targeted attitudes	302	67,500	.290	.271	.219	.246	.295	.001	.540	91.98
Performance : all knowledge, skills, & abilities[b]	1,110	1,182,197	.258	.266	.232	.252	.280	-.027	.560	97.68
Performance : all psychological characteristics	2,680	735,327	.205	.204	.171	.197	.210	-.003	.411	88.58

Performance : all objective person characteristics	1,259	863,852	.123	.106	.107	.100	.112	-.023	.234	87.58
Turnover : attitudes[c]	461	191,285	.163	.168	.135	.156	.181	.007	.330	87.49
Turnover : organization attitudes	219	89,590	.147	.172	.139	.154	.191	.006	.339	88.03
Turnover : job attitudes[c]	196	77,059	.183	.182	.123	.165	.200	.037	.328	84.44
Turnover : psychological characteristics	254	109,364	.129	.105	.093	.093	.116	.002	.207	73.94
Turnover : objective person characteristics[d]	312	293,428	.101	.068	.064	.061	.075	-.003	.139	74.33

[a]*k*: number of samples; N: number of observations; unwt: unweighted; wt: sample size-weighted; SD*r*: standard deviation of *r*; CONF: confidence interval; CRED: credibility interval; I^2: index of heterogeneity.
[b]30 influential cases (i.e., N > 10,000) were removed from this analysis.
[c]5 influential cases removed from this analysis.
[d]11 influential cases removed from this analysis.

Table 4
Sample Sizes Needed to Achieve .80 Power as a Function of Variable Relation Type

	Effect Size Distribution Percentile								
Relation Type	**20th**	**25th**	**33rd**	**40th**	**50th**	**60th**	**67th**	**75th**	**80th**
(All effect sizes)	3137	1599	966	542	304	175	123	79	58
Attitudes : attitudes	782	462	269	175	105	65	52	36	29
Organization attitudes : Job attitudes	542	346	215	146	91	61	52	40	31
Organization attitudes : People attitudes	398	269	146	97	65	46	40	31	26
Job attitudes : People attitudes	1224	782	398	269	146	97	84	61	55
Attitudes : intentions	646	398	240	159	105	69	52	38	33
Attitudes : behaviors	2178	1599	782	542	304	193	134	91	65
Intentions : behaviors	2178	966	646	462	240	146	113	74	61
Performance : attitudes	2178	1224	782	462	304	175	123	79	61
Performance : organization attitudes	1599	1224	782	462	304	215	175	105	84
Performance : job attitudes	2178	1599	966	646	304	193	123	74	58
Performance : people attitudes	966	542	346	240	134	84	58	44	34
Performance:knowledge, skills, & abilities	1224	782	398	304	159	113	79	52	38
Performance:psychological characteristics	2178	1224	782	542	304	193	146	97	74
Performance:objective person characteristics	8718	4903	3137	2178	1224	646	462	304	215
Turnover : attitudes	3137	2178	1224	782	462	269	193	134	113
Turnover : organization attitudes	4903	2178	1599	966	646	398	269	146	134
Turnover : job attitudes	2178	1599	646	462	304	193	159	113	97
Turnover : psychological characteristics	4903	3137	2178	1224	782	462	304	215	193
Turnover : objective person characteristics	19619	8718	3137	2178	1599	966	646	398	304

FIGURE 1

Abbreviated Hierarchical Variable Taxonomy

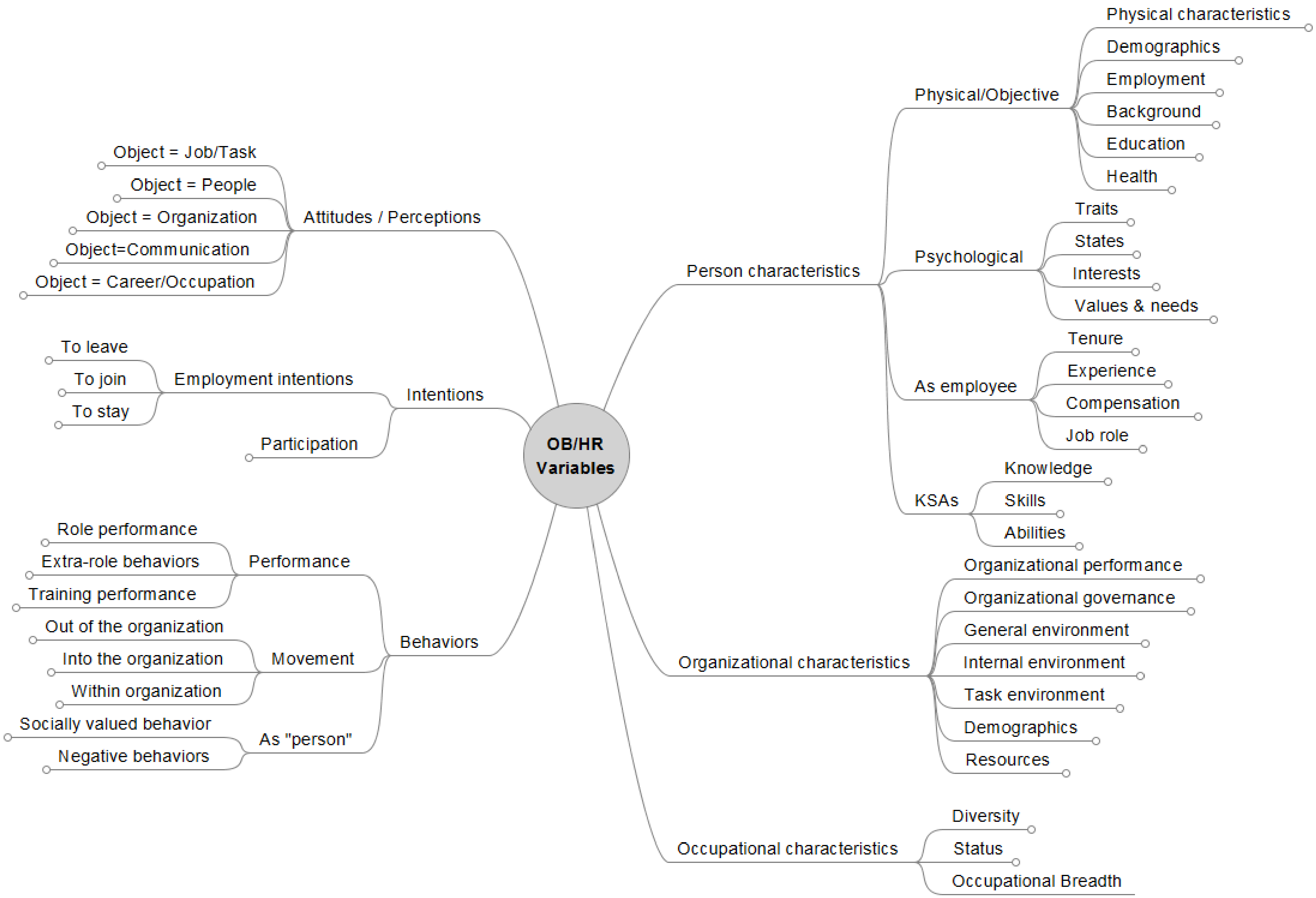

FIGURE 2

Ranges for Classification as a "Moderate" Effect Size as a Function of Source