Journal of Applied Psychology
2015, Vol. 100, No. 2, 431–449

0021-9010/15/$12.00 http://dx.doi.org/10.1037/a0038047

Correlational Effect Size Benchmarks

Frank A. Bosco
Virginia Commonwealth University

Herman Aguinis
Indiana University

Kulraj Singh
South Dakota State University

James G. Field
Virginia Commonwealth University

Charles A. Pierce
University of Memphis

Effect size information is essential for the scientific enterprise and plays an increasingly central role in the scientific process. We extracted 147,328 correlations and developed a hierarchical taxonomy of variables reported in *Journal of Applied Psychology* and *Personnel Psychology* from 1980 to 2010 to produce empirical effect size benchmarks at the omnibus level, for 20 common research domains, and for an even finer grained level of generality. Results indicate that the usual interpretation and classification of effect sizes as small, medium, and large bear almost no resemblance to findings in the field, because distributions of effect sizes exhibit tertile partitions at values approximately one-half to one-third those intuited by Cohen (1988). Our results offer information that can be used for research planning and design purposes, such as producing better informed non-nil hypotheses and estimating statistical power and planning sample size accordingly. We also offer information useful for understanding the relative importance of the effect sizes found in a particular study in relationship to others and which research domains have advanced more or less, given that larger effect sizes indicate a better understanding of a phenomenon. Also, our study offers information about research domains for which the investigation of moderating effects may be more fruitful and provide information that is likely to facilitate the implementation of Bayesian analysis. Finally, our study offers information that practitioners can use to evaluate the relative effectiveness of various types of interventions.

Keywords: effect size, statistical analysis, null hypothesis testing, big data

Effect size (ES) estimates provide an indication of relation strength (i.e., magnitude), are essential for the scientific enterprise, and are "almost always necessary" to report in primary studies (American Psychological Association, 2010, p. 34; Kelley & Preacher, 2012). Moreover, ES information plays an increasingly central role in the scientific process, informing study design (e.g., a priori power analysis; hypothesis development), statistical analysis (e.g., meta-analysis; Bayesian techniques; Kruschke, Aguinis, & Joo, 2012), and the assessment of scientific progress (Cohen, 1988; Cumming, 2012; Grissom & Kim, 2012; Ozer, 1985), as well as practical significance (Aguinis et al., 2010; Brooks, Dalal, & Nolan, 2014). It should come as no surprise that Cohen (1988) stated, "a moment's thought suggests that [ES] is, after all, what science is all about" (p. 532).

ES awareness has risen partly due to the increased popularity of Cohen's (1962, 1988) benchmarks for classifying correlations of $|r| = .1, .3, .5$ as small, medium, and large, respectively. However, Cohen's (1962, 1988) benchmarks are "controversial" (Ellis, 2010b, p. 40), and their generalizability to findings in applied psychology is currently unknown. In addition, important knowledge regarding effect sizes has been derived from meta-analyses in particular domains such as personnel selection (e.g., Roth, BeVier, Bobko, Switzer, & Tyler, 2001), conceptual analyses of reasons why validity coefficients seem to reach a ceiling in many research domains (e.g., Cascio & Aguinis, 2008b), fluctuations in effect sizes across different measures of similar constructs (e.g., Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995), and the literature on convergent and discriminant validity, which makes researchers sensitive to the relative highs and lows of effect size estimates (e.g., Carlson & Herdman, 2012). In spite of these advancements, there is a need for applied psychologists to know

This article was published Online First October 13, 2014.

Frank A. Bosco, Department of Management, School of Business, Virginia Commonwealth University; Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University; Kulraj Singh, Department of Economics, College of Agriculture and Biological Sciences, South Dakota State University; James G. Field, Department of Management, School of Business, Virginia Commonwealth University; Charles A. Pierce, Department of Management, Fogelman College of Business and Economics, University of Memphis.

A previous version of this article was presented at the August 2013 meetings of the Academy of Management, Orlando, FL.

We thank Allison Gabriel for her comments on a previous draft.

Correspondence concerning this article should be addressed to Frank A. Bosco, Department of Management, School of Business, Virginia Commonwealth University, Richmond, VA 23284-4000. E-mail: fabosco@vcu.edu

more about the overall level of scientific success of our field and how the level of success varies across studied phenomena (e.g., turnover vs. performance) and general variable types (e.g., intention vs. behavior).

Cohen's (1962) ES benchmarks were intuited from results reported in the 1960 volume of *Journal of Abnormal and Social Psychology*: $|r| = .2$, .4, and .6 as small, moderate (i.e., medium), and large effect sizes, respectively. The benchmarks were later revised ($|r| = .1$, .3, .5; Cohen, 1988), yet still based on a nonempirical approach (Aguinis & Harden, 2009). Importantly, although Cohen's (1988) benchmarks have become the norm (Hill, Bloom, Black, & Lipsey, 2008) and are widely adopted (e.g., as input to power analysis; Aguinis & Harden, 2009), some researchers have argued that the prescribed minimum cutoff values (i.e., $|r| \geq .30$ for a medium effect; see Ellis, 2010b; Ferguson, 2009) are unrealistically high (e.g., Hemphill, 2003). Importantly, Cohen's (1988) benchmarks came with no generalizability guarantee. In fact, Cohen (1988) noted that a researcher who finds that "what is here defined as 'large' is too small (or too large) to meet what his area of behavioral science would consider appropriate standards is *urged* to make more suitable operational definitions" (p. 79; italics added). However, whether Cohen's (1988) guidelines—or any single, omnibus set of guidelines—can depict the corpus of research findings in applied psychology is currently unknown. Moreover, how can researchers "make more suitable operational definitions"?

The purpose of this study is to present solutions to several key challenges associated with ES benchmarks in applied psychology. We apply an innovative data collection protocol that allowed us to empirically define a single, omnibus benchmark and a diverse set of context-specific ES benchmarks for relation types commonly investigated in applied psychology. Also, we have made our database available so that researchers can use it to extract benchmarks at different levels of generality. Overall, we present new, empirically based benchmarks and describe benefits of their adoption for stages of the scientific process and scientific progress.

Our article is organized as follows. First, we describe the ubiquitous role of ES use and interpretation throughout the research process. Second, we describe benchmark refinement efforts from other areas of psychological and social science research that provide examples of benefits brought by refined field level (i.e., omnibus) and finer grained benchmarks. Third, we report the results of a study including approximately 150,000 correlational effect size estimates published in *Journal of Applied Psychology* and *Personnel Psychology* from 1980 to 2010. We classified each effect size in terms of its relation type and provide a refined set of omnibus ES benchmarks, as well as 20 benchmarks for coarse and fine-grained relation types. Also, we make our database available and illustrate how it can be used to derive effect size benchmarks at several different levels of generality—including narrower levels that have been reported in some published meta-analyses. We discuss applications of effect size benchmarks for better-informed non-nil hypotheses, study design (e.g., a priori power analysis), and the interpretation of results. In addition, we discuss future applications of our findings, including the facilitation of Bayesian statistical techniques and the identification of research domains where searches for moderating effects are likely to be more fruitful. Finally, we describe implications for practice, focusing on the interpretation of intervention effectiveness within and across research domains.

State of Effect Size Awareness

Although Cohen's guidelines for interpreting effect sizes have been adopted widely, a brief review of their use in applied psychology reveals inconsistent interpretation. As noted earlier, Cohen (1988) defined small, moderate (i.e., medium), and large $|r|$ as "about" .10, .30, and .50, respectively (p. 185). What remains uncertain, however, is what exactly *about* represents. Ellis's (2010b) interpretation treats Cohen's values as minimum cutoffs that, for example, define the range of medium ES as $.30 \leq |r| < .50$. Others classify effect sizes in terms of their surrounding anchors (e.g., $r = .39$ as medium to large; Rosnow & Rosenthal, 2003). Another interpretation is that Cohen's values represent range centroids. For example, Rhoades and Eisenberger's (2002) interpretation of Cohen's (1988) medium ES range, $.24 \leq |r| < .36$, is centered at .30. Still other approaches appear to combine ranges and cutoffs (Rudolph, Wells, Weller, & Baltes, 2009). Ferguson (2009) involved practical significance in a set of benchmarks, defining $r = .20$ as the minimum practically significant value, with minimum cutoffs for moderate and large effect sizes at $r = .50$ and $r = .80$, respectively. Figure 1 includes a graphical depiction of moderate ES range according to each interpretation. In short, there is lack of clarity regarding collective effect size awareness (i.e., interpretation guidelines) and also a lack of clarity regarding the actual distribution of ES magnitudes in the field.

Tailored, updated ES benchmarks have been developed in the areas of international management (Ellis, 2010a), psychological treatment (Hemphill, 2003), and neuropsychology (Zakzanis, 2001). As an example, Hemphill's (2003) benchmarks defined a medium ES between $|r| = .18$ and .30, a substantial departure from Cohen's (1988) benchmarks. Importantly, psychological treatment researchers now benefit from a frame of reference that allows for better informed contrasts (e.g., between particular treatments) and an indication of the overall effectiveness of their area of inquiry (i.e., psychological treatment). Researchers in education (Hill et al., 2008) have developed even finer grained benchmarks reflecting particular intervention and sample types. Indeed, as Hemphill (2003) stated, "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" (p. 79).

We acknowledge that the use of context-specific benchmarks may gloss over differences across fields in the ability to model outcomes (we also address this issue in the Limitations section). On the other hand, journal editors have expressed the need for contextualized benchmarks for the purpose of evaluating substantive significance (Ellis, 2010a), and such benchmarks also play a critical role in science-based practice (Hill et al., 2008). For example, Wilkinson and the APA Task Force on Statistical Inference (1999) noted that "we must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses" (p. 599).

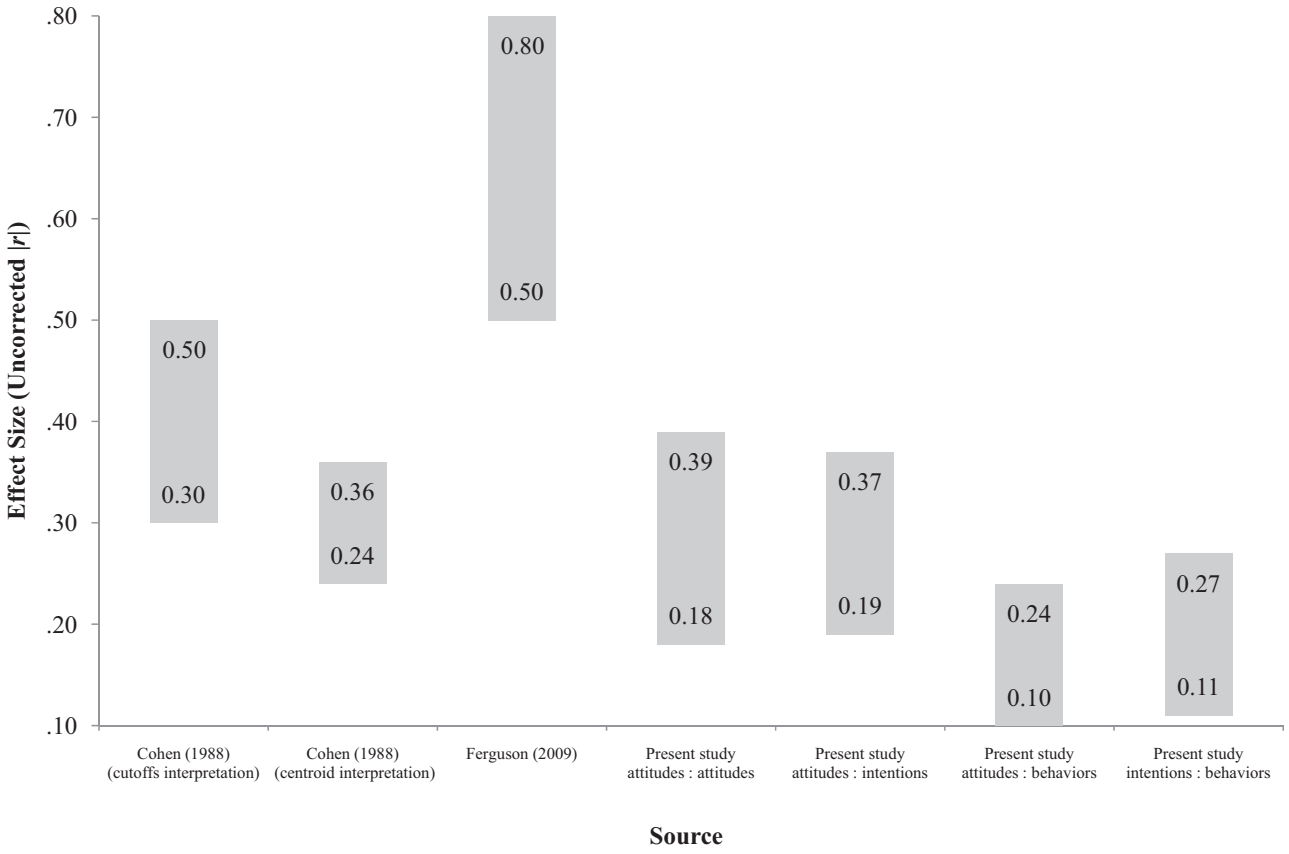

Figure 1. Ranges for classification as a "medium" or "moderate" effect size, as a function of source.

Present Study

The present study provides a large-scale analysis of applied psychology research from a database of 147,328 correlational effect sizes (*r*s) published in *Journal of Applied Psychology* or *Personnel Psychology* from 1980 to 2010. From analyses of the effect sizes 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 applied psychology? To answer this question, we present the most comprehensive set of field-level, omnibus ES benchmarks and contrast them with existing benchmarks. As a second research question, we ask: Are common bivariate relation "types" associated with different ES distributions? To this end, we provide ES benchmarks for 20 common relation types in applied psychology research (e.g., attitude–intention vs. attitude–behavior relations) and describe how more refined benchmarks can better inform several research processes. In addition, we illustrate how researchers can zoom in on the broader types of relationships to obtain finer grained correlational effect sizes at a desired level of generality. Taken together, we provide an empirically based understanding of ES distributions in applied psychology research—broadly and in particular contexts—that can be used to assess scientific progress, estimate practical significance, and inform many important decisions regarding study design and data-analytic techniques such as a priori power analysis and Bayesian inference.

Method

Database

We collected all correlation coefficients reported in primary study tables of *Journal of Applied Psychology* (JAP) and *Personnel Psychology* (PPsych) articles from 1980 to 2010. The data presented in this article are part of a broader data collection effort. We excluded meta-analyses and articles whose purpose was to reanalyze an earlier data set (i.e., we included only original, empirical articles reporting at least one table or matrix of correlation coefficients). A total of 1,660 unique articles containing 147,328 effect sizes and their respective sample sizes are included in the database. We conducted analyses at the ES unit of analysis, which we transformed into absolute values prior to analysis (the list of articles is available from the authors upon request, and the database is available at http://www.frankbosco.com/data).

To code for variable type, the first author created an initial taxonomy based on existing typologies in applied psychology research (Cascio & Aguinis, 2008a; Crampton & Wagner, 1994). After extracting variable names from a subset of the articles' correlation tables, we followed the approach by Aguinis, Pierce, Bosco, and Muslin (2009) and refined the taxonomy through several rounds of error checks and discussions among the first, third, and fourth authors. As an example of the hierarchical structure, attitudes are categorized in terms of their respective targets

Table 1
Examples of Variable Types Used to Classify 147,328 Correlational Effect Size Estimates Reported in Journal of Applied Psychology *and* Personnel Psychology, *1980–2010*

Variable	Example
People attitudes	Supervisor satisfaction; coworker satisfaction; leader–member exchange
Job attitudes	Job satisfaction; autonomy perceptions; pay satisfaction
Organization attitudes	Organizational commitment; perceived organizational support; procedural justice
Intentions	Turnover intention; intent to accept a job offer; intent to participate in development
Behavior	Performance; absenteeism; turnover
Performance	In-role performance; extra-role performance; training performance
KSAs	Job knowledge; decision-making skills; general mental ability
Psychological characteristics	Traits (e.g., conscientiousness; core self-evaluation) and states (e.g., stress; burnout)
Objective person characteristics	Age; gender; tenure
Movement	Voluntary turnover; job choice; involuntary turnover

Note. KSAs = knowledge, skills, and abilities.

(e.g., attitudes toward the job; toward people; toward the organization). Similarly, behaviors are categorized in terms of their major types (e.g., performance; employee movement). Specificity increases at finer levels of the taxonomy. The taxonomy is comprehensive and covers all major topics in industrial–organizational psychology, organizational behavior, and human resource management textbooks. In total, the taxonomy arranges 4,869 nodes (i.e., variable names or category names) into 10 first-level nodes (e.g., behavior; attitude; intention), which then branch to a mean of 5.2 second-level nodes (e.g., behavior: performance; behavior: movement: turnover), third-level nodes, and so forth.

Examples of major variable types from within four of the 10 first-level nodes are shown in Table 1. For illustrative purposes, a highly abbreviated version of the classification taxonomy is shown in Figure 2. Figure 2 includes only six of the 10 first-level nodes included in the unabbreviated taxonomy and an even much smaller subset of the 4,869 nodes included in the entire taxonomy.

Relations were coded in three levels of abstraction: coarse, fine, and extra fine. In terms of frequency, four common, coarse relation types emerged: attitudes–attitudes, attitudes–intentions, attitudes–behaviors, and intentions–behaviors. In addition, we identified four fine bivariate relation types with performance behavior (attitudes–performance; knowledge, skills, and abilities–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 movement behavior, such as voluntary turnover (attitudes–movement; psychological characteristics–movement; objective person characteristics–movement), and two extra fine relation types for the attitudes–movement relation (organization attitudes–movement; job attitudes–movement). Although the sample size (i.e., the number of effect size estimates) for people attitudes–movement was smaller than 200, we include estimates for this relation type.

In contrast to existing typologies (e.g., Cascio & Aguinis, 2008a), our taxonomy of variables presents a taxonomic display concerning what variables actually represent rather than how they are used. As an example, although personality traits are categorized as a predictor of employee performance in existing typologies (e.g., Cascio & Aguinis, 2008a), they are also used as a predictor of employee turnover and other organizationally relevant outcomes (Zimmerman, 2008). In contrast, in the present taxonomy, personality traits are categorized more broadly under the first-level node: person characteristics. In addition, we treat the

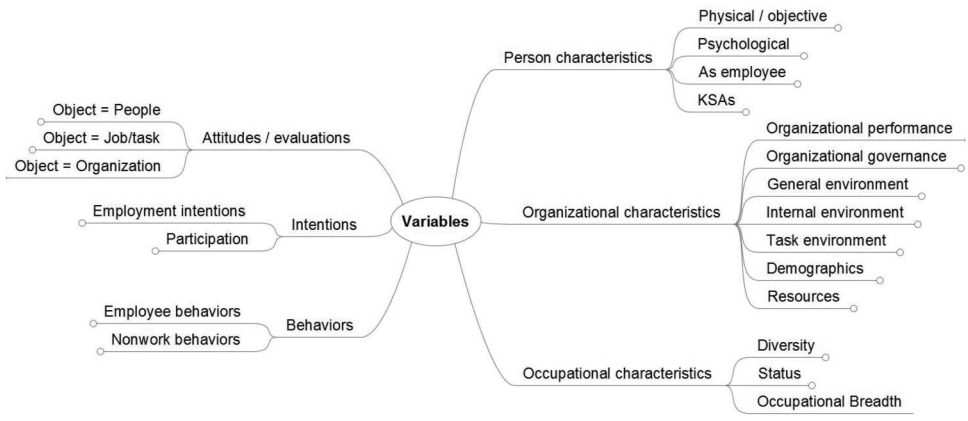

Figure 2. Abbreviated hierarchical variable taxonomy used to classify 147,328 correlational effect size estimates reported in *Journal of Applied Psychology* and *Personnel Psychology* from 1980 to 2010 (the total number of nodes is 4,869).

attitude concept broadly in our taxonomy, as the attitudes literature has for decades (Fazio, Sanbonmatsu, Powell, & Kardes, 1986). Generally, attitudes represent cognitive and/or affective evaluations of a given target. An attitude target may be virtually any *thing*—an individual, an event, an organizational policy, and even a parking spot. Although one could make the argument that supervisory ratings of performance are themselves attitudes (i.e., where a supervisor maintains an attitude toward the attitude target: an employee's output), we nonetheless classify supervisory ratings of performance as an indicator of performance behavior.

The present analyses are based on a database of 1,660 unique articles containing 25,891 variables and 147,328 effect sizes. Thus, articles contain a mean of 88.75 effect sizes, or roughly the equivalent of one 14 × 14 correlation matrix. Many articles contain more than one correlation matrix (i.e., to present findings for multiple samples or studies). As mentioned earlier, following other reviews and syntheses of correlational effect sizes (e.g., Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011), we conducted our analyses at the ES level because dependence is unlikely to threaten the validity of our inferences as it might in survey research (i.e., Kish, 1965, pp. 257–263). As noted by Glass, McGaw, and Smith (1981), "The data set to be [meta-]analyzed will invariably contain complicated patterns of statistical dependence . . . ∵ each study is likely to yield more than one finding . . . The simple (but risky) solution . . . is to regard each finding as independent of the others" (p. 200). Although originally labeled as "risky," the Glass et al. (1981) recommendation has been supported by Monte Carlo simulation results. Specifically, Tracz, Elmore, and Pohlmann (1992) noted that "even a cursory review of published meta-analyses reveals that the assumption of independence is, in fact, seldom met" (p. 881). Reassuringly, however, results of Tracz et al.'s Monte Carlo simulations provided evidence that "nonindependence of the data does not affect the estimation of the population parameter, rho" (p. 883). Thus, their conclusion was that "proceeding under the assumption of independence is not so risky as previously thought . . . ∵ combining the statistics from nonindependent data in a correlational meta-analysis does not have an adverse effect on the results" (Tracz et al. 1992, p. 886).

Coding Process and Agreement

The third and fourth authors coded all variables in the data set according to the taxonomy. Thus, for each of the 25,891 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) is located in the taxonomy as a fifth-level node (i.e., attitudes → attitudes toward people → attitudes toward supervisors/mentors → exchange → LMX). Coders used a combination of exact letter string matching with the taxonomy's node text and decision making to code each variable. Infrequent variables (e.g., prejudicial attitudes against West Germans) were coded as miscellaneous by assigning a broad classification node (e.g., attitudes toward people).

After the 25,891 variables were coded according to the taxonomy, we used database tools in Microsoft Excel to create the list of 147,328 effect sizes with taxonomy node codes for each variable in the pair. Thus, if a given correlation matrix contained 14 variables, only the 14 variables' taxonomic assignments required manual coding. From these 14 codes, a total of 91 bivariate relation code pairs were produced and linked to the ES and sample size information in the database using range lookup formulas.

To assess coder agreement, articles were randomly selected until each coder had independently coded 301 effect sizes. Then, we assessed 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). The two coders agreed on 278 (92.4%) of the 301 assignments.

Results

Omnibus Field-Level Benchmarks

Our first research question asks to what extent existing ES benchmarks reflect the extant applied psychology literature. To this end, we describe the omnibus distribution of the 147,328 effect sizes 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 bare-bones meta-analytic estimates for each ES distribution.

As shown in Table 2, the distribution of 147,328 effect sizes exhibits a median ES of $|r| = .16$ and is split into thirds (i.e., upper and lower boundaries for medium ES) at $|r| = .09$ and .26. Our observed medium ES range is thus similar to Dalton, Aguinis, Dalton, Bosco, and Pierce's (2012) ES distributions split into thirds at $|r| = .10$ and .22 (published effect sizes) and $|r| = .11$ and .28 (nonpublished effect sizes), but substantially different (i.e., non-overlapping medium ES range) when compared to Cohen's (1988) benchmarks by any interpretation (see Figure 1). In addition, as shown in Table 2, we observed values of $|r| = .05$, .07, .12, .21, .32, and .36 for the 20th, 25th, 40th, 60th, 75th, and 80th percentiles of the omnibus ES distribution, respectively. Importantly, Cohen's (1988) benchmarks for small, medium, and large ESs (i.e., $|r| = .10$, .30, .50) correspond to approximately the 33rd, 73rd and 90th percentiles, respectively, of our distribution of 147,328 effect sizes. Finally, as shown in Table 2, effect sizes in the center tertile of our omnibus distribution are classified as medium by Cohen's (1988) benchmarks in only 8.2% of cases (centroid interpretation) or 0% of the cases (cutoffs interpretation).

As a second analytic approach to summarizing the distribution of the 147,328 ESs, we conducted a bare-bones meta-analysis (i.e., correcting for the biasing effect of sampling error only). As shown in Table 3, our analysis revealed a mean ES that is small by both interpretations of Cohen's (1988) standards, ($|r| = .222$; 95% CI = .221, .223; $k = 147{,}328$; N = 325,218,877). The unweighted mean ES revealed a similar value, $|r| = .219$. As might be expected with a large, diverse collection of effect sizes, 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 data set ($I^2 = 98.97$), and the 80% credibility interval (–.03, .48) includes zero (Hunter & Schmidt, 2004).

The median ES value reported above, $|r| = .16$, is smaller than the mean meta-analytically derived ES, $|r| = .22$, indicating that the distribution of effect sizes is positively skewed ($\text{skew}_{|r|} = 1.27$;

Table 2
Effect Size Distribution Percentiles for Broad Relation Types

Relation type	*k*	*N*	ES distribution percentile									Overlap with Cohen's medium ES range[a]	
			20th	25th	33rd	40th	50th	60th	67th	75th	80th	Cutoffs[b]	Centroid[c]
(All effect sizes)	147,328	225	.05	.07	.09	.12	.16	.21	.26	.32	.36	0.00%	8.21%
Attitudes: attitudes	14,493	202	.10	.13	.18	.22	.28	.34	.39	.45	.50	40.26%	56.52%
Organization attitudes: Job attitudes	1,263	240	.14	.16	.21	.25	.31	.36	.40	.45	.49	55.58%	61.96%
Organization attitudes: People attitudes	644	277	.15	.18	.24	.28	.34	.39	.43	.48	.51	70.45%	61.36%
Job attitudes: People attitudes	783	196	.10	.13	.18	.21	.26	.30	.35	.40	.43	25.82%	62.18%
Attitudes: intentions	1,717	237	.12	.15	.19	.23	.27	.33	.37	.42	.47	37.46%	66.61%
Attitudes: behaviors	7,958	220	.06	.07	.10	.12	.16	.20	.24	.29	.33	0.00%	0.00%
Intentions: behaviors	535	233	.07	.09	.11	.14	.19	.24	.27	.32	.33	0.00%	15.34%
Performance: attitudes	3,224	190	.07	.08	.11	.14	.17	.22	.26	.31	.36	0.00%	9.30%
Performance: organization attitudes	615	213	.07	.08	.10	.13	.16	.19	.22	.27	.30	0.00%	0.00%
Performance: job attitudes	1,271	188	.06	.08	.10	.13	.17	.22	.26	.32	.36	0.00%	9.85%
Performance: people attitudes	575	192	.08	.10	.13	.16	.22	.27	.32	.39	.43	6.77%	38.02%
Performance: knowledge, skills, & abilities	1,385	202	.08	.10	.13	.16	.21	.26	.31	.36	.40	4.80%	32.99%
Performance: psychological characteristics	3,135	158	.06	.07	.10	.12	.16	.20	.23	.28	.31	0.00%	0.00%
Performance: objective person characteristics	1,395	200	.03	.04	.05	.07	.09	.11	.14	.17	.20	0.00%	0.00%
Movement: attitudes	866	309	.05	.07	.09	.11	.14	.18	.21	.25	.28	0.00%	0.00%
Movement: org. attitudes	200	309	.07	.08	.10	.13	.14	.19	.23	.27	.30	0.00%	0.00%
Movement: job attitudes	295	312	.06	.07	.09	.11	.13	.16	.18	.22	.25	0.00%	0.00%
Movement: people attitudes	44	266	.06	.06	.09	.09	.12	.21	.23	.31	.37	0.00%	0.00%
Movement: psychological characteristics	288	216	.04	.05	.07	.08	.11	.13	.17	.20	.23	0.00%	0.00%
Movement: objective person characteristics	461	293	.02	.03	.04	.05	.07	.09	.11	.14	.16	0.00%	0.00%

Note. Percentiles show the distribution divided into 2, 3, 4, and 5 equal partitions. ES = effect size; *k* = number of effect sizes; *N* = median sample size.
[a] Represents the percentage of ES that are classified medium by Cohen's (1988) benchmarks and also in the center tertile of present study's ES distributions. We omit comparisons with Ferguson's (2009) benchmarks. [b] Based on Ellis's (2010b) interpretation of Cohen's (1988) medium ES range (i.e., $.30 \leq |r| < .50$). [c] Based on Rhoades and Eisenberger's (2002) interpretation of Cohen's (1988) medium ES range (i.e., $.24 \leq |r| < .36$).

$skew_r = 0.33$). A positively skewed ES distribution was expected because, as noted by Cohen (1988), large effect sizes are relatively rare in social science research. In fact, although our study is based on absolute value effect sizes, the distribution of raw ES values for one of applied psychology's largest meta-analyses on a single topic (Judge, Thoresen, Bono, & Patton, 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 Effect Size Benchmarks

Our second research question asks whether different major relation types exhibit distinct ES distributions. As described earlier, distinct within-discipline 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. We identified 20 common, broad bivariate relation types. Several categorizations contain overlapping ES sets (e.g., performance and turnover are subsets of behavior in our taxonomy). We followed 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 with comparisons to Cohen's (1988) benchmark interpretations (see Table 2). In addition, we provide bare-bones meta-analytic values (i.e., corrected for sampling error) for each bivariate relation type (see Table 3). Finally, as an additional set of results, Table 4 presents sample sizes needed to achieve .80 power a priori (Cohen, 1988) for each relation type. Although we present values to achieve a power level of .80, we present the inputs needed to estimate any level of power.

As shown in Table 2, there is substantial variance in ES distribution parameters across the 20 bivariate relation types. Specifically, the four coarse relation types provide definitions of medium effect sizes with partitions at $|r| = .18$ and .39 (attitudes–attitudes), $|r| = .19$ and .37 (attitudes–intentions), $|r| = .10$ and .24 (attitudes–behaviors), and $|r| = .11$ and .27 (intentions–behaviors). Thus, for relations involving behaviors, ES values greater than roughly $|r| = .25$ exist in the upper tertile 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 6.50% variance explained (i.e., uncorrected $|r| = .255$) when predicting behavior represents a large ES in that context, but the corresponding value for a large ES among nonbehavioral relations (i.e., attitudes–attitudes; attitudes–intentions) is 14.44% (i.e., uncorrected $|r| = .380$). Thus, in many contexts, Cohen's (1988) benchmarks are nonapplicable by a factor of two or more.

Values for the three fine relation types with employee performance are also shown in Table 2. Results reveal medium ES boundaries at $|r| = .13$ and .31 (knowledge, skills, and abilities–performance), $|r| = .10$ and .23 (psychological characteristics–

Table 3
Bare-Bones Meta-Analytic Estimates for Broad Relation Types

Relation type	*k*	*N*	unwt mean *r*	wt mean *r*	SD_r	95% CI lower	95% CI upper	80 % Cred lower	80 % Cred upper	I^2
(All effect sizes)	147,328	325,218,877	.219	.222	.200	.221	.223	−.033	.477	98.97
Attitudes: attitudes	14,493	6,675,710	.310	.290	.207	.286	.293	.030	.549	95.73
Organization attitudes: Job attitudes	1,263	611,778	.319	.371	.206	.360	.383	.112	.631	96.38
Organization attitudes: People attitudes	644	328,597	.342	.330	.195	.315	.346	.085	.576	95.93
Job attitudes: People attitudes	783	311,296	.285	.256	.176	.244	.269	.039	.473	92.87
Attitudes: intentions	1,717	804,084	.297	.283	.190	.274	.292	.046	.520	94.99
Attitudes: behaviors	7,958	3,845,993	.207	.180	.184	.176	.184	−.049	.409	94.28
Intentions: behaviors	535	302,123	.218	.158	.148	.146	.171	−.024	.340	92.32
Performance: all attitudes	3,224	915,077	.223	.203	.173	.197	.209	−.006	.413	89.17
Performance: organization-targeted attitudes	615	177,338	.195	.196	.162	.183	.209	.002	.390	87.69
Performance: job-targeted attitudes	1,271	326,771	.221	.196	.167	.187	.205	−.003	.395	86.99
Performance: people-targeted attitudes	575	147,112	.268	.251	.211	.234	.268	−.009	.510	92.26
Performance: all knowledge, skills, & abilities	1,385	1,327,369	.255	.381	.303	.365	.397	−.005	.768	99.17
Performance: all psychological characteristics	3,135	799,506	.202	.196	.171	.190	.202	−.009	.400	87.48
Performance: all objective person characteristics	1,395	668,815	.127	.089	.102	.084	.095	−.028	.206	80.17
Movement: attitudes	866	946,866	.172	.103	.120	.095	.111	−.046	.251	93.75
Movement: organization attitudes	200	89,723	.190	.204	.143	.184	.224	.031	.378	90.00
Movement: job attitudes	295	684,872	.154	.072	.090	.062	.082	−.041	.185	94.78
Movement: people attitudes	44	19,849	.201	.176	.158	.129	.222	−.018	.369	91.81
Movement: psychological characteristics	288	130,423	.143	.114	.104	.102	.126	−.005	.234	80.19
Movement: objective person characteristics	461	4,866,496	.107	.026	.040	.023	.030	−.024	.076	94.19

Note. *k* = number of effect sizes; *N* = number of observations; unwt = unweighted; wt = sample size weighted; *SDr* = standard deviation of *r*; CI = confidence interval; Cred = credibility interval; I^2 = index of heterogeneity not accounted for by sampling error.

performance), $|r|$ = .05 and .14 (objective person characteristics–performance), and $|r|$ = .11 and .26 (attitudes–performance). We acknowledge that a global average including what can be an ill-defined population of attitudes may not be informative. Accordingly, Table 2 also shows that three extra fine relations within the attitudes–performance relation type reveal medium ES partitions at $|r|$ = .10 and .22 (organization attitudes–performance), $|r|$ = .10 and .26 (job attitudes–performance), and $|r|$ = .13 and .32 (people attitudes–performance). Thus, although broad in nature, 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.

Results regarding the three fine relation types with employee movement behavior are also presented in Table 2. Medium ES boundaries are at $|r|$ = .07 and .17 (psychological characteristics–movement), $|r|$ = .04 and .11 (objective person characteristics–movement), and $|r|$ = .09 and .21 (attitudes–movement). In addition, two extra fine relation types for the attitudes–movement relation type revealed medium ES partitions at $|r|$ = .10 and .23 (organization attitudes–movement) and $|r|$ = .09 and .18 (job attitudes–movement). Finally, although we located only 44 effect sizes for the people attitudes–movement relation type, we observed medium tertile partitions for this category at $|r|$ = .09 and .23. Thus, our findings reveal that employee movement behavior is predicted relatively poorly compared to performance behavior. In addition, broadly, such relations with employee movement behavior larger than roughly $|r|$ = .20 exist within the top third of the ES distribution in that context (i.e., large effect sizes).

As shown in Table 2, center tertiles for coarse nonbehavioral relations exhibit roughly 60% overlap with Cohen's (1988) benchmarks (centroid interpretation) or 40% overlap with Cohen's benchmarks (cutoffs interpretation). Indeed, the centroids interpretation of Cohen's (1988) benchmarks places $|r|$ = .30 at the *center* of the medium ES range. The corresponding centroid values for the present analyses are $|r|$ = .28 (attitudes–attitudes) and $|r|$ = .27 (attitudes–intentions). Thus, as suggested by Cohen (1988), medium effect sizes for these two particular bivariate relation types are about .30. However, medium effect sizes for coarse relation types involving behaviors (i.e., attitude–behavior; intention–behavior) are substantially smaller. For coarse behavioral relations, the overlap comparing Cohen's (1988) centroid-based medium ES range and the present analyses ranges from 0% to 15% (0% for the cutoffs interpretation).

Table 3 shows bare bones meta-analytic results for the omnibus and category-specific relation type ES distributions. Similar to the 50th percentile values displayed in Table 2, Cohen's (1988) medium ES centroid ($|r|$ = .30) seems to depict nonbehavioral relations (i.e., attitude–attitude; attitude–intention) but not behavioral relations (i.e., attitude–behavior; intention–behavior). In addition, as expected, the omnibus and 20 category-specific meta-analytic estimates indicate high levels of between-study variability not due to sampling error (i.e., $I^2 > .75$ benchmark; Higgins & Thompson, 2002). Stated differently: As expected, potential for moderation detection is high among all sets of effect sizes. Indeed, given the coarseness and scope of our analyses, one would expect high degrees of heterogeneity. However, among the 21 I^2 values (M = 91.86; SD = 5.17), two relation types presented with I^2 values

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

Relation type	Effect size distribution percentile								
	20th	25th	33rd	40th	50th	60th	67th	75th	80th
(All effect sizes)	3,137	1,599	966	542	304	175	113	74	58
Attitudes: attitudes	782	462	239	159	97	65	49	36	29
Organization attitudes: Job attitudes	398	304	175	123	79	58	46	36	30
Organization attitudes: People attitudes	346	239	133	97	65	49	40	31	27
Job attitudes: People attitudes	782	462	239	175	113	84	61	46	40
Attitudes: intentions	542	346	215	146	105	69	55	42	33
Attitudes: behaviors	2,177	1,599	782	542	304	193	133	91	69
Intentions: behaviors	1,599	966	646	398	215	133	105	74	69
Performance: attitudes	1,599	1,224	646	398	269	159	113	79	58
Performance: organization attitudes	1,599	1,224	782	462	304	215	159	105	84
Performance: job attitudes	2,177	1,224	782	462	269	159	113	74	58
Performance: people attitudes	1,224	782	462	304	159	105	74	49	40
Performance: knowledge, skills, & abilities	1,224	782	462	304	175	113	79	58	46
Performance: psychological characteristics	2,177	1,599	782	542	304	193	146	97	79
Performance: obj. person characteristics	8,718	4,903	3,137	1,599	966	646	398	269	193
Movement: attitudes	3,137	1,599	966	646	398	239	175	123	97
Movement: organization attitudes	1,599	1,224	782	462	398	215	146	105	84
Movement: job attitudes	2,177	1,599	966	646	462	304	239	159	123
Movement: people attitudes	2,177	2,177	966	966	542	175	146	79	55
Movement: psychological characteristics	4,903	3,137	1,599	1,224	646	462	269	193	146
Movement: objective person characteristics	19,620	8,718	4,903	3,137	1,599	966	646	398	304

Note. Sample size values are based on the two-tailed exact test for bivariate normal correlations using G*Power (Faul et al., 2009).

more than 2 *SD* units below the mean: objective person characteristics–performance (I^2 = 80.17) and psychological characteristics–employee movement (I^2 = 80.19).

As shown in Table 4, sample sizes required to achieve .80 a priori power (Faul, Erdfelder, Buchner, & Lang, 2009) vary considerably across content domain. Indeed, using our coarse benchmarks, sample sizes required to achieve .80 power for a 50th percentile ES vary between 97 and 150 (for nonbehavioral relations), and between 215 and 304 (for behavioral relations). In addition, in all cases where relation types are comparable, employee movement (e.g., turnover) studies require larger sample sizes than studies related to individual performance.

Finally, our database can be used to extract effect sizes at an even more fine-grained level of generality. For example, assume there is an interest in zooming in on the coarse "Behaviors" category which, as shown in Figure 2, is one of the first-level categories. Figure 3 includes an illustrative subset of nodes that branch out of the broad "Behaviors" category. The total number of nodes under "Behaviors" in the database is 1,163, but, for illustrative purposes, Figure 3 shows only 48 of these nodes. Assume that there also is an interest in focusing on another one of the six broad categories shown in Figure 2: "Attitudes/evaluations." For illustrative purposes, Figure 4 shows a graphic representation of a subset of 56 of the 1,103 nodes under this broad category.

By zooming in on each of the broad categories, we are able to subsequently extract effect sizes at many different levels of generality. For example, assume we would like to know the size of relationships between the broad category "attitudes" with behaviors that range in the level of generality from the most general level (i.e., all behaviors combined) to finer and finer levels down to "Facet/task subjective role performance," which is a seventh-level node (see Figure 3). Table 5 includes these results, which are quite informative. For example, the 50th percentile for the relationship between attitudes and goal performance is .43, whereas the same percentile for the relationship between attitudes and job search behaviors is .17. In addition, I^2 values shown in Table 5 are informative regarding which types of relationship are more likely to lead to fruitful moderation research. For example, the I^2 value for the relationship between attitudes and absenteeism/tardiness is only 48.55, whereas the value for the relationship between attitudes and group/team performance is 92.48, suggesting the presence of moderators in the latter but not necessarily the former relationship.

As a second illustration of the use of our database to examine relationships at different levels of generality, consider now the possibility of focusing on the broad category "Attitudes/evaluations" (see Figure 4). Table 6 shows an illustrative subset of such relations. For example, the 50th percentile correlation between behaviors and organizational image attitudes/evaluations is .26, whereas the 50th percentile correlation between behaviors and compensation attitudes/evaluations is only .12. In addition, Table 6 shows variability regarding I^2 values suggesting that moderation research is not likely to be fruitful regarding, for example, behavioral relations with the identity core characteristic of the job characteristics model (i.e., I^2 = 51.31) are less likely to reveal moderation compared to those with feedback core characteristic (i.e., I^2 = 90.47). In short, the illustrative Tables 5–6 and Figures 3–4 show that the database can be used for various levels of precision and, in some cases, a higher level of precision than some published meta-analyses.

In sum, results indicate that commonly used, existing ES benchmarks are not appropriately tailored to the applied psychology research context. In addition, results indicate that empirical benchmarks for ES magnitude vary as a function of bivariate relation type.

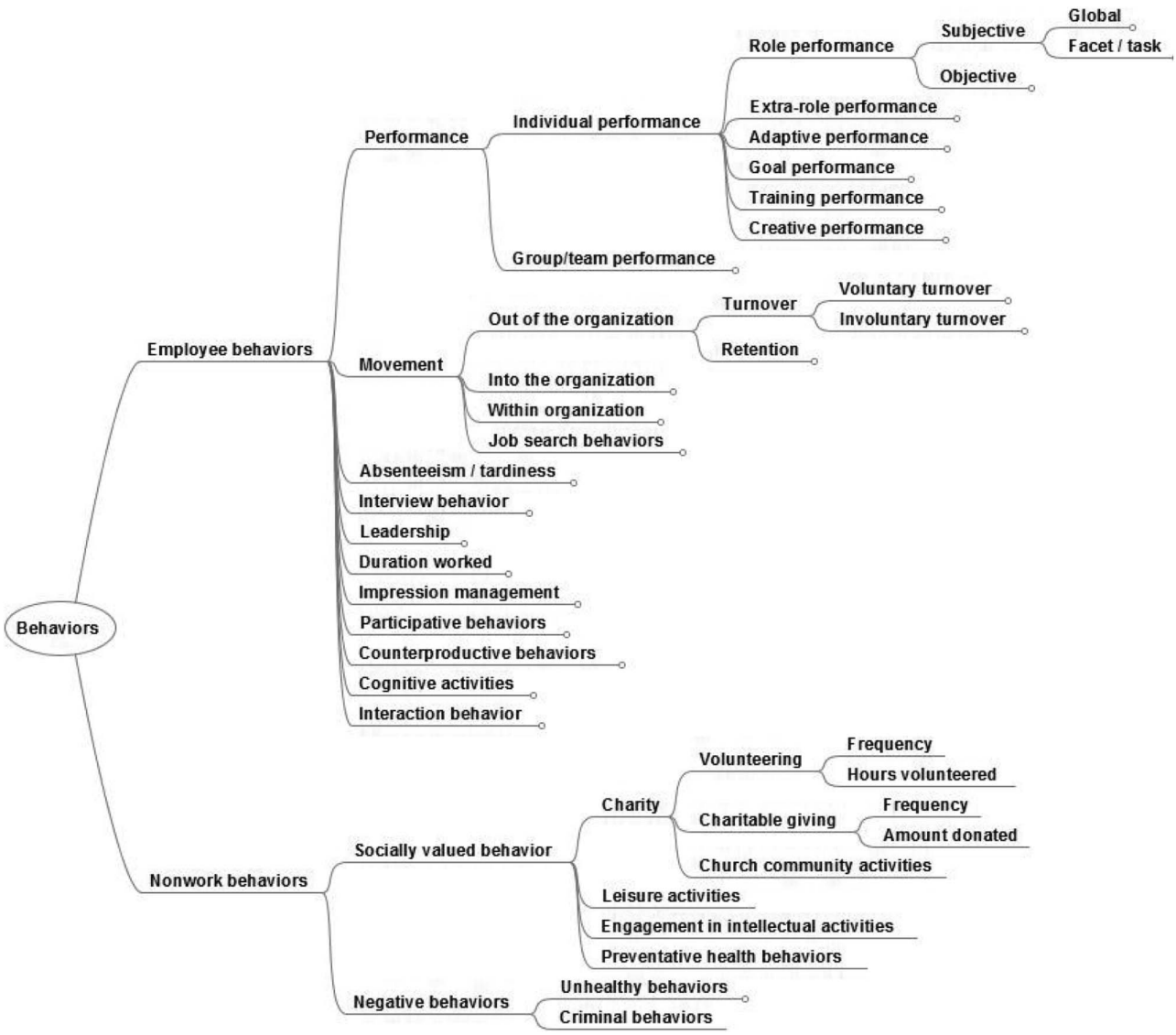

Figure 3. Abbreviated hierarchical variable taxonomy used to classify behavioral variables reported in *Journal of Applied Psychology* and *Personnel Psychology* from 1980 to 2010 (the total number of nodes is 1,163).

Discussion

As Hill et al. (2008) 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" (p. 177). 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 1, results of the present study indicate that the ES benchmark generalizability concern originally raised by Cohen (1988) himself and echoed by others (e.g., Hemphill, 2003; Hill et al., 2008) is well-founded.

Our first research question addressed the extent to which Cohen's ES benchmarks reflect the omnibus distribution of findings in applied psychology research. Our results indicate that none of the existing benchmark operationalizations described previously fit findings in applied psychology. Specifically, the existing cutoffs-based guidelines classify 0% of our omnibus center-tertile effect sizes as medium in size. Centroid-based guidelines perform only slightly better, classifying 8.21% of the omnibus center tertile as medium in size. Thus, at the omnibus and highest level of generality and aggregation, many applied psychology research results have been interpreted and classified with an effect size rubric that bears almost no resemblance to findings in the field.

Our second research question addressed the extent to which benchmarks vary across bivariate relation type (e.g., attitude–intention vs. attitude–behavior). Results indicate substantial variance in empirical definitions of medium ES across relation types, and thus one single benchmark will not suffice (see Table 2). At the broadest level of our taxonomy, relations involving behaviors (i.e., attitude–behavior; intention-behavior) are substantially smaller than others (i.e., attitude–attitude; attitude–intention). Indeed, the greatest degree of classifica-

Figure 4. Abbreviated hierarchical variable taxonomy used to classify attitudinal variables reported in *Journal of Applied Psychology* and *Personnel Psychology* from 1980 to 2010 (the total number of nodes is 1,103).

tion overlap comparing the present coarse benchmarks to the centroid interpretation of Cohen's (1988) benchmarks is 15% for relations involving behaviors (i.e., attitude = behavior; intention–behavior) and approximately 60% for others (i.e., attitude–attitude; attitude–intention). For heuristic purposes, our results indicate that medium effect sizes involving behaviors (i.e., attitudes–behaviors; intentions–behaviors) are between roughly $|r| = .10$ and .25. In contrast, for relations not involving behaviors (e.g., attitudes–attitudes; attitudes–intentions), medium effect sizes are between roughly $|r| = .20$ and .40. Our study is not the first to highlight effect size fluctuations across research domains, constructs, and measures (e.g., Bommer et al., 1995). But a unique value-added contribution of our study and database is that our results show the little overlap between these two center tertiles and the ability to extract effect sizes ranging in their level of generality, resulting in a number of important implications for theory and research as well as practice.

Implications for Theory and Research

Non-nil predictions. The present results provide useful information that can be used for making theoretical advancements in the future. Specifically, it has been argued that an effective way to promote theoretical advancement is to increase theoretical precision by deriving non-nil predictions, such that theories predict the presence of a nonzero effect rather than the mere absence of a zero effect (Edwards & Berry, 2010; Meehl, 1990). Although non-nil predictions can be found in the natural sciences, such as physics and chemistry, they are rare in applied psychology research. In

Table 5
Effect Size Distribution Percentiles and Bare-Bones Meta-Analytic Estimates for Illustrative Relations Between Attitudes/Evaluations at the Broad Level of Generality and Behaviors at Broad and Finer Levels of Generality (Figure 3)

Relation type: Attitudes/evaluations with . . .	*k*	Med. *N*	Mean *N*	25th	33rd	50th	67th	75th	Unwt mean |*r*|	*N*-wtd mean |*r*|	SD_r	80% Cred	I^2
Behaviors	7,958	220	483	.07	.10	.16	.24	.29	.21	.18	.18	(−.05, .41)	94.28
Employee behaviors	7,736	217	479	.07	.10	.16	.24	.29	.21	.18	.18	(−.05, .41)	94.28
Performance	3,224	190	284	.08	.11	.17	.26	.31	.22	.20	.17	(−.01, .41)	89.17
Individual performance	2,737	192	276	.09	.11	.18	.26	.31	.22	.21	.17	(.00, .42)	88.98
Role performance	1,797	192	275	.08	.10	.17	.26	.31	.22	.22	.18	(−.01, .44)	90.12
Subjective	1,205	185	276	.10	.12	.19	.29	.35	.25	.25	.19	(.01, .49)	91.59
Global	604	161	253	.10	.12	.19	.28	.33	.24	.24	.18	(.01, .46)	89.65
Facet/task	555	221	299	.10	.13	.21	.31	.37	.26	.27	.20	(.02, .52)	92.93
Objective	515	193	292	.05	.07	.12	.18	.21	.16	.14	.13	(−.01, .29)	80.77
Extra-role performance	605	199	241	.11	.14	.20	.28	.32	.23	.22	.15	(.04, .41)	84.38
Goal performance	58	62	69	.13	.24	.43	.63	.72	.46	.46	.31	(.08, .85)	91.01
Training performance	167	182	402	.06	.08	.14	.22	.29	.19	.13	.13	(−.03, .28)	85.53
Creative Performance	48	285	220	.09	.09	.13	.20	.26	.19	.15	.15	(−.02, .33)	81.37
Group/team performance	147	92	257	.09	.12	.19	.37	.39	.26	.26	.21	(−.01, .52)	92.48
Movement	866	309	1,093	.07	.09	.14	.21	.25	.17	.10	.12	(−.05, .25)	93.75
Out of the organization	346	306	936	.07	.08	.12	.18	.21	.15	.10	.10	(−.03, .22)	89.96
Turnover	270	306	1,045	.06	.07	.10	.14	.16	.11	.07	.07	(.00, .15)	78.49
Voluntary turnover	80	327	2,672	.05	.06	.09	.11	.14	.10	.06	.06	(−.01, .12)	87.96
Involuntary turnover	12	785	699	.06	.07	.10	.10	.11	.08	.08	.04	(.07, .09)	12.25
Retention	73	861	562	.20	.24	.29	.32	.34	.27	.29	.09	(.19, .40)	81.80
Into the organization	17	354	320	.16	.22	.31	.37	.39	.32	.32	.18	(.09, .54)	92.94
Within organization	120	320	1,647	.04	.07	.15	.21	.25	.17	.05	.11	(−.09, .18)	94.78
Job search behaviors	334	278	405	.07	.10	.17	.25	.28	.19	.20	.14	(.02, .37)	89.03
Absenteeism/tardiness	628	271	295	.05	.06	.11	.15	.17	.12	.11	.08	(.04, .18)	48.55
Interview behavior	79	266	387	.08	.10	.14	.23	.33	.23	.19	.22	(−.08, .46)	94.87
Leadership	689	416	693	.14	.17	.25	.33	.38	.28	.39	.25	(.08, .70)	98.28
Duration worked	143	196	788	.04	.05	.08	.14	.20	.14	.08	.09	(−.02, .19)	85.12
Impression management	16	64	108	.16	.17	.20	.28	.64	.34	.28	.26	(−.03, .59)	88.65
Participative behaviors	429	248	388	.09	.13	.21	.32	.38	.25	.24	.19	(.00, .48)	93.84
Counterproductive behaviors	546	374	1,120	.06	.08	.14	.20	.23	.17	.13	.12	(−.01, .28)	93.79
Cognitive activities	293	171	205	.07	.10	.16	.26	.31	.22	.18	.16	(.00, .37)	81.75
Interaction behavior	563	182	211	.07	.09	.15	.24	.29	.19	.20	.17	(.00, .40)	84.39
Non-employee behavior	212	469	643	.04	.06	.10	.15	.20	.15	.12	.15	(−.07, .30)	93.26
Socially valued behavior	43	242	264	.07	.09	.12	.18	.19	.15	.13	.10	(.03, .24)	65.69
Negative behaviors	127	470	797	.05	.06	.09	.14	.18	.14	.10	.13	(−.06, .25)	92.46
Unhealthy behaviors	120	470	816	.05	.06	.09	.13	.16	.14	.09	.13	(−.06, .25)	92.38

Note. Percentiles show the distribution divided into 2, 3, and 4 equal partitions. *k* = number of effect sizes; Med. = median; unwt = unweighted; *N*-wtd = sample size weighted; SD_r = standard deviation of *r*; Cred = credibility interval; I^2 = index of heterogeneity not accounted for by sampling error.

fact, predictions stated as point estimates are often difficult to justify. Our results offer ranges of values, akin to the "good-enough" belt advocated by Serlin and Lapsley (1985) and referred to by others (e.g., Edwards & Christian, 2014). Specifically, results summarized in Tables 2–6 can be used to derive non-nil predictions. For example, a study investigating a relationship between attitudes and behavior would state a non-nil hypothesis that the expected effect will be at least |*r*| = .16 rather than zero (as is used within a null hypothesis significance testing [NHST] framework).

Study design. Our results also have implications for several research process stages. During the research design stage, an anticipated ES is necessary to conduct a priori power analysis (a process that informs the data collection phase). While existing meta-analytic estimates and/or direct replications represent suitable sources, specifying the targeted effect size is the "most difficult part of power analysis" (Cohen, 1992, p. 156). Our results offer an alternative approach for nascent research areas: anticipated ES specification based on broad relation types. Indeed, power analysis should rely on the most context-specific ES benchmarks available (Hill et al., 2008). However, when an existing estimate 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 take 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 applied psychology research context, the use of which could lead to upwardly biased ES forecasts and thus underpowered studies (Maxwell, 2004).

As Cohen (1988, 1992) noted, power analysis is essential for research planning and aids in the reduction of Type II errors. At the field level, the median sample size for effect sizes reported in *JAP* from 1995 to 2008 is 173 (Shen et al., 2011). At a sample size of 173, only anticipated effect sizes greater than |*r*| = .21 would have achieved statistical power greater than .80. Our results indicate 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 power = .80. Using the two median values just described

Table 6
Effect Size Distribution Percentiles and Bare-Bones Meta-Analytic Estimates for Illustrative Relations Between Behaviors at the Broad Level of Generality and Attitudes/Evaluations at Broad and Finer Levels of Generality (Figure 4)

Relation type: Behaviors with . . .	*k*	Med. *N*	Mean *N*	25th	33rd	50th	67th	75th	Unwt mean |*r*|	*N*-wtd mean |*r*|	SD_r	80% Cred	I^2
Attitudes/evaluations	7,958	220	483	.07	.10	.16	.24	.29	.21	.18	.18	(−.05, .41)	94.28
Object = job/task	2,972	224	674	.07	.09	.15	.22	.27	.19	.15	.17	(−.07, .37)	95.36
Job characteristics	1,308	244	670	.07	.09	.14	.21	.25	.18	.13	.13	(−.02, .29)	91.27
JCM	484	270	291	.07	.09	.15	.24	.29	.19	.19	.17	(−.01, .40)	89.02
Identity	39	332	340	.09	.11	.15	.16	.17	.14	.15	.08	(.08, .21)	51.31
Significance	54	332	367	.03	.05	.07	.10	.14	.11	.08	.09	(−.01, .17)	66.97
Autonomy	241	260	272	.11	.13	.21	.30	.36	.24	.25	.18	(.04, .46)	89.52
Feedback	122	279	297	.04	.06	.12	.19	.28	.17	.17	.18	(−.05, .39)	90.47
Stressors	371	193	349	.06	.08	.14	.20	.24	.17	.16	.13	(.00, .32)	84.71
Job scope	42	332	341	.12	.15	.17	.18	.21	.18	.18	.09	(.09, .27)	65.06
Knowledge characteristics	166	332	3,146	.07	.09	.14	.22	.28	.17	.10	.07	(.01, .19)	94.39
Roles	130	240	280	.04	.07	.09	.15	.19	.15	.22	.23	(−.07, .50)	93.84
General job affect	927	226	772	.07	.10	.16	.23	.27	.19	.16	.23	(−.12, .45)	97.63
Compensation	201	271	687	.04	.06	.12	.16	.19	.14	.13	.11	(−.01, .26)	89.19
Performance appraisal system	175	178	240	.07	.09	.15	.26	.34	.22	.22	.19	(−.01, .45)	89.56
Goals	130	62	103	.10	.13	.23	.41	.53	.34	.27	.27	(−.05, .60)	88.69
Object = organization	1,456	241	407	.08	.10	.16	.23	.28	.20	.24	.20	(−.02, .49)	94.73
Org policies/procedures	378	225	344	.08	.12	.18	.24	.28	.21	.20	.18	(−.02, .42)	91.82
Justice	289	253	354	.08	.12	.18	.23	.27	.20	.19	.17	(−.02, .40)	90.95
Interpersonal justice	23	231	724	.12	.14	.20	.29	.31	.23	.10	.14	(−.08, .28)	93.78
Interactional justice	32	229	250	.08	.13	.20	.27	.30	.22	.20	.16	(.02, .38)	85.20
Distributive justice	85	270	326	.06	.07	.13	.19	.23	.17	.17	.16	(−.01, .36)	88.34
Procedural justice	129	225	347	.09	.13	.19	.23	.28	.21	.22	.17	(.01, .43)	91.04
Employee–organization relationship	851	233	428	.08	.10	.15	.22	.27	.20	.27	.22	(−.01, .54)	95.72
Perceived organizational performance	34	90	404	.06	.07	.10	.15	.19	.14	.11	.10	(.00, .22)	76.21
Embeddedness	51	310	295	.09	.10	.14	.24	.26	.17	.17	.12	(.04, .30)	77.17
Organizational image	36	612	616	.12	.14	.26	.31	.40	.26	.29	.16	(.09, .48)	94.70
Satisfaction towards organization	25	785	743	.05	.05	.08	.09	.11	.10	.11	.12	(−.05, .26)	91.86
Object = people	1,338	199	290	.07	.10	.17	.27	.33	.23	.22	.19	(−.02, .46)	91.74
Super/managers/leaders	626	237	342	.08	.11	.18	.27	.34	.23	.24	.20	(.00, .48)	93.27
Supervisor support	100	248	363	.07	.09	.12	.18	.23	.16	.17	.13	(.02, .32)	84.13
Supervisor trust	22	124	185	.10	.11	.16	.22	.33	.23	.32	.23	(.05, .60)	91.94
Abusive supervision	29	216	248	.17	.17	.19	.26	.28	.23	.24	.15	(.06, .42)	84.94
Supervisor satisfaction	120	259	442	.09	.13	.19	.36	.47	.28	.34	.25	(.03, .66)	97.19
Coworkers	302	142	234	.08	.12	.19	.29	.37	.25	.20	.19	(−.03, .44)	89.45

Note. Percentiles show the distribution divided into 2, 3, and 4 equal partitions. JCM = job characteristics model; *k* = number of effect sizes; Med. = median; unwt = unweighted; N-wtd = sample size weighted; SD_r = standard deviation of *r*; Cred = credibility interval; I^2 = index of heterogeneity not accounted for by sampling error.

(i.e., $N = 173$; Shen et al., 2011) and $|r| = .16$ from the present analyses, the median statistical power in applied psychology research based on correlation coefficients published in JAP and PPsych from 1980 to 2000 is .56. Alternatively, using the median sample size observed in the present study ($N = 224$) and the median ES ($|r| = .16$), the median finding is associated with power = .67. Thus, applied psychology research still appears to suffer from insufficient statistical power, and we hope that refined benchmarks will reduce this problem by providing a more realistic estimate of sample size needed to achieve statistical power in context.

Interpretation of results. Table 7 presents detailed examples of reported effect sizes in JAP articles with (re)interpretations and recommendations in relation to ES benchmarks. As an example, an uncorrected meta-analytic estimate on the hiring expectancies—job choice relation, $r = .16$, ($k = 6$; $N = 720$; 95% CI = .09, .24) was described by the authors as "small" (Chapman, Uggerslev, Carroll, Piasentin, & Jones, 2005, p. 935), a classification that could serve to shift attention away from this relation. However, while the ES is classified as small by Cohen's benchmarks, our findings indicate that it exists at roughly the 50th percentile among attitude/evaluation–behavior relations. Indeed, the best of 12 unique predictors of job choice reported in Chapman et al.'s (2005) meta-analysis presented with $r = .17$. Thus, in this context, hiring expectancy is worthy of attention for those interested in explaining, to at least some degree, job choice. Clear, low-cost implications could flow from this reinterpretation, including suggested modifications to communication style and frequency with job candidates.

Our results help shed light on important questions for the entire field. As an example, applied psychologists have had relatively more success predicting employee performance than employee movement (e.g., turnover). Indeed, performance is better predicted than movement among all six parallel comparisons in the present study (see Table 2). One explanation is that employee movement measures (e.g., turnover behavior) are dichotomous and thus require corrected effect sizes. Whatever the cause, it is important that researchers develop an ES awareness in which, for example, it is

Table 7
Examples of Obtained Effect Size, Classification, and Reinterpretation Based on Effect Size Benchmarks

Source	Relation	Effect size	Benchmark classifications	Authors' interpretation	Alternative interpretation
Arthur et al. (2006)	P–O fit: employee performance	r = .12 Uncorrected meta-analytic r (k = 36; N = 5,377; 95% CI [.08, .17])	•Cohen (1988; cutoffs): Small •Cohen (1988; centroid): Small •Ferguson (2009): Not practically significant •Updated benchmarks: Medium (37th percentile in the performance–organization attitudes/evaluations context)	"Thus, the job performance effect was small" (Cohen, 1992, p. 793). "We recommend that organizations should exercise caution when using P–O fit to make employment-related decisions (e.g., selection) in the absence of local validation studies or until new research refutes the findings obtained here" (p. 797).	Practitioners should note that the median ES for relations of this broad type (i.e., attitudes/evaluations toward the organization–performance) is r = .16. For organizations seeking to employ organizational–attitudinal evaluations (e.g., during selection to increase performance), we recommend seeking out other attitudes/evaluations providing r = .16 (50th percentile) or larger. Especially efficacious predictors in this context have ESs greater than r = .22 (67th percentile) or .27 (75th percentile). To achieve r = .30 (80th percentile) is noteworthy.
Chapman et al. (2005)	Hiring expectancies: Job choice	r = .16 Uncorrected meta-analytic r (k = 6; N = 720; 95% CI = .09, .24)	•Cohen (1988; cutoffs): Small •Cohen (1988; centroid): Small •Ferguson (2009): Not practically significant •Updated benchmarks: Medium (55th percentile among movement–attitudes/evaluations ESs; 50th among the broader the attitudes–behaviors context)	"All predictors of job choice had either small effects or were not significant" (p. 935).	Among 12 estimates for predictors of job choice, person–job fit is the largest (r = .17), but nonsignificant. Perceived hiring expectancies (r = .16) significantly predict job choice, at a median ES level (50th percentile) given the broad attitude/evaluation–behavior context.

(*table continues*)

Table 7 (*continued*)

Source	Relation	Effect size	Benchmark classifications	Authors' interpretation	Alternative interpretation
					Although many attitudes/evaluations fail to predict job choice behavior, perceived hiring expectancy seems to be an exception. Based on these findings, we recommend that practitioners consider strategies to manage hiring expectancies. This is a low-cost strategy that might mitigate the loss of key candidates to competitors and foster the formation of positive attitudes toward the organization and its employees.
Parker et al. (2006)	Coworker trust: proactive work behavior	$r = .15$ (Primary study; uncorrected r; $234 \leq N \leq 282$)	•Cohen (1988; cutoffs): Small •Cohen (1988; centroid): Small •Ferguson (2009): Not practically significant •Updated benchmarks: Medium (37th percentile among performance–people attitudes/evaluations ESs)	"Coworker trust was found to be an antecedent of proactive work behavior, albeit having a relatively small effect" (p. 647).	Among other attitudes toward people-performance relations, co-worker trust is a medium ES (37th percentile) below the median. Importantly, among other broad attitude types (i.e., attitudes toward the organization or job) attitudes toward people predict performance best. Organizations seeking to assess employee attitudes toward others (e.g., coworkers, as part of climate assessment) to effect a change in performance level should consider similar attitudes that provide at least $r = .22$ (50th percentile). In this context, $r = .32$ (67th percentile) and $r = .39$ (75th percentile) are the largest effect sizes.

Table 7 (*continued*)

Source	Relation	Effect size	Benchmark classifications	Authors' interpretation	Alternative interpretation
Rhoades and Eisenberger (2002)	POS: extra-role performance toward the organization	r = .24 (Uncorrected meta-analytic r; k = 8; N = 2,079)	•Cohen (1988; cutoffs): Small •Cohen (1988; centroid): Medium •Ferguson (2009): .04 larger than minimum cutoff for practical significance •Updated benchmarks: Large (70th percentile among performance–organization attitudes/evaluations ESs)	"The relationship between POS and extrarole performance directed to the organization was medium sized" (p. 710).	The median level of predictive success for relations of this broad type (i.e., performance–attitudes/evaluations toward the organization) is r = .16. The present observe value, r = .24, is classified as a large ES in this context (i.e., 70th percentile). Predictive success at the 80th percentile (r = .30) is especially laudable. Practitioners seeking to effect desired performance outcomes should give special consideration to POS and, if space is limited in yearly attitudinal surveys, should consider replacing other similar organization attitudes/evaluations with this one.

known that success in explaining variance in employee performance is roughly double that of employee movement. In short, knowledge regarding effect sizes across areas leads to questions such as the ones above, which point to fruitful areas for future research.

As additional implications, our findings indicate that psychological characteristics (e.g., personality variables) predict performance and movement to a greater degree than objective person characteristics (e.g., demographic variables). The present findings also indicate that attitudes and intentions perform relatively similarly when predicting behaviors and that attitude–attitude and attitude–intention relations are among the strongest substantive ESs observed in applied psychology. These findings can serve as the basis for an omnibus assessment of a long-standing, meta-theoretic view on the attitude–intention–behavior mediated model and distinctiveness of attitudes and intentions (Fishbein & Ajzen, 1975). In short, ES benchmarks facilitate the development of a framework wherein research results may be interpreted.

An additional implication of our results is that knowledge of effect sizes 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 described earlier, the present results indicate that, overall, researchers tend to have more success predicting employee performance compared to employee turnover. We submit that for scientific progress to continue to advance, we must first realize where progress is not being made. Although the solution to a slow rate of scientific progress is beyond the scope of the present study, we propose that field-level analyses can indicate areas where more research would be beneficial (cf. Colquitt & Zapata-Phelan, 2007).

Identification of moderating effects. Our results also offer insights into future research that would benefit particularly from assessing contingent (i.e., interactive) relationships. Interactive relationships lay at the heart of theories regarding person–environment fit, individual performance, differential prediction and validity in selection research, and any theory that considers outcomes to be a result of the joint influence of two or more variables (e.g., Grizzle, Zablah, Brown, Mowen, & Lee, 2009; Mathieu, Aguinis, Culpepper, & Chen, 2012; Wallace, Edwards, Arnold, Frazier, & Finch, 2009; Yu, 2009). However, researchers often lament lack of success in finding support for such contingent relationships (e.g., Aguinis, Beaty, Boik, & Pierce, 2005; Mathieu et al., 2012). Results summarized in Tables 3, 5, and 6 offer useful information in terms of the variability of bivariate relationships across research domains. Recall that I^2 describes the percentage of variation across studies that is due to heterogeneity rather than chance. Thus, values for I^2 allow us to rank order research domains in terms of the presence of potential moderator variables (i.e., contingent factors) that explain why a particular bivariate relationship varies across primary-level studies. For example, it is more likely that moderators will be found in future research addressing knowledge, skills, and abilities as predictors of performance compared to psychological characteristics as predictors of employee performance. Thus, results presented in our tables, combined with appropriate theoretical rationale, can serve to guide future empirical research assessing moderating effects.

Bayesian analysis. An additional implication of our results concerns the application of Bayesian techniques, the use of which seems to be creating a revolution in fields ranging from genetics to marketing (Kruschke et al., 2012). Bayesian approaches are becoming popular because they do not rely on null hypothesis significance testing, which is known to have several problems (e.g., Cortina & Landis, 2011). Specifically, what researchers want to know are the parameter values that are credible, given the observed data. In particular, researchers may want to know the viability of a null hypothesis (e.g., zero correlation between two variables) given the data, $p(\text{Ho}|\text{D})$; Kruschke et al., 2012). However, traditional methods based on NHST tell us the probability of obtaining the data in hand, or more extreme unobserved data, if the null hypothesis were true, $p(\text{D}|\text{Ho})$ (Aguinis et al., 2010). Unfortunately, $p(\text{Ho}|\text{D}) \neq p(\text{D}|\text{Ho})$. As noted by Cohen (1994), a test of statistical significance "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997).

Although Bayesian approaches are appealing, an important challenge, which is often seen as the Achilles' heel of Bayesian analysis, is the need to specify a prior distribution of effect sizes. Indeed, as noted by Kruschke et al. (2012), "the prior distribution is not capricious and must be explicitly reasonable to a skeptical scientific audience" (p. 728). Our results regarding the distribution of effect sizes across various research domains provide empirically based, explicit, and reasonable anchors that make Bayesian analysis in applied psychology more feasible in the future.

Implications for Practice

Our results have implications for the communication of findings and the estimation of practical significance. Imagine that a human resources practitioner encounters a recent uncorrected meta-analytic estimate of the general mental ability (GMA)–performance relation ($r = .28$; cf. Schmidt, Shaffer, & Oh, 2008). In addition, the manager considers 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 likely clear to the manager that personality traits explain relatively little variance in performance compared to GMA, armed with Cohen's (1988) cutoff-based benchmark lens the manager might conclude that all of these predictors exhibit small effects (i.e., $|r| < .30$; Cohen, 1988; a different interpretation is reached with the centroid-based interpretation of Cohen's benchmarks).

In contrast, by providing practitioners with the present empirically derived benchmarks (i.e., omnibus medium ESs between $|r| = .09$ and .26), we submit that three of the Big Five personality traits (agreeableness, extraversion, and openness to experience) present with small effect sizes (i.e., lower tertile; $|r| < .09$). In addition, for conscientiousness ($r = .14$) and emotional stability ($r = .09$), the minimum qualification for classification as a medium ES is met. However, according to the present benchmarks, GMA presents with a large (i.e., upper tertile) ES ($r = .28$; Schmidt et al., 2008), a classification compatible with the statement that, among the options available, "intelligence is the best predictor of job performance" (Ree & Earles, 1992, p. 86). We

present four additional examples of benchmarks at varying levels of generality to show their influence on substantive conclusions and recommendations for practice in Table 7.

Limitations and Future Research Directions

Our study relies on the assumption of the usefulness of relativistic benchmarks. The use of context-specific benchmarks is likely to gloss over important differences across fields in the ability to model outcomes. As noted by an anonymous reviewer, it seems problematic that the same relationship would be referred to as a small effect size in one domain and medium in another. Taken to a logical extreme, every meta-analysis would produce a medium effect size, because a pool of studies in the meta-analysis would define the relevant reference population on that topic. So, a key question is the following: What is the optimal level of generality for the benchmarks? Our approach to answering this question is threefold. First, we reported benchmarks for a high level of generality. Specifically, Table 2 shows benchmarks based on the entire database of correlational effect sizes and also benchmarks for broad areas and research domains. This information offers the highest level of generality. Second, Table 2 also includes benchmarks at the fine and extra fine levels of generality. Moreover, Tables 5–6 and Figures 3–4 illustrate the possibility of zooming in on the broader types of relationships summarized in Table 2 to obtain finer-grained correlational effect sizes at a desired level of generality. Third, we make our database available online (http://www.frankbosco.com/data) which allows researchers to extract benchmarks for various levels of generality that are useful for different purposes and objectives. In short, our approach allows readers to explore the database to obtain estimates at a given desired level of generality.

A second limitation is that we have only summarized effect sizes found in tables of two (albeit prestigious and influential) applied psychology journals from 1980 to 2010. It remains possible that effect sizes found in other journals from other points in time might reveal different distribution parameters. We note, however, that our analyses are based on the largest individual-level ES database in the field; although they included a wider range of journal sources, other similar content analyses have involved substantially smaller data sets (e.g., Aguinis et al., 2011; Dalton et al., 2012; Shen et al., 2011). In addition, it remains possible that by including only top-tier journals, our ES estimates are upwardly biased (e.g., due to different types of publication bias; Kepes & McDaniel, 2013).

As another limitation, our analyses present a coarse overview of relation types and include thousands of ESs that were not hypothesized per se but were included in correlation matrices nonetheless. As an example, correlations between the demographic variables age and sex are included in our single omnibus ES benchmark. Relation type heterogeneity decreases, however, as one considers the finer level benchmarks. As such, the present ES distribution parameters could be downwardly biased. However, as described earlier, the compromise position we present necessitates some degree of taxonomic coarseness.

Although obtained effect sizes feed estimates of practical significance, information on effect sizes alone does not suffice in terms of communicating a study's impact on practice (Aguinis et al., 2010; Brooks et al., 2014). As a general guideline, Ferguson (2009) suggested that a "recommended minimum [ES] representing a 'practically' significant effect for social science data" (p. 533) is $r = .20$. Importantly, however, Ferguson noted that "scholars are cautioned that effect size interpretation should be context specific" (p. 533) and that the "guidelines are suggested as minimum cutoffs, not guarantees that effect sizes exceeding those cutoffs are meaningful" (p. 536). The present study provides a starting point, in terms of context-specific ES distributions, for this exercise.

As additional future research directions, researchers should consider the development of multiple ES benchmarks. The precision of benchmark output would be enhanced by applying multiple benchmarks. Indeed, as noted by Hill et al. (2008, p. 177), "it is often useful to use multiple benchmarks when assessing the observed impacts of an intervention." As an example, a context-specific benchmark for, say, predictive versus concurrent validation study design could be combined with benchmarks for, say, psychological characteristics–performance. While the precise combinatorial method is beyond the scope of the present study, 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.

Finally, our study addresses the correlation coefficient, which is the effect size metric most frequently used and reported in the applied psychology literature (e.g., Aguinis et al., 2011). The reason for the pervasiveness of r as an indicator of effect size is that the majority of research published in *JAP* and *PPsych* is non-experimental. However, other types of effect sizes such as d (Cohen, 1988), which are typically used in the context of experimental designs, are also reported in the applied psychology literature. Accordingly, because effects resulting from the use of experimental designs are likely to be larger than those resulting from passive observation designs, an interesting avenue for future research is an investigation of noncorrelational effect size benchmarks, for example, based on d.

Conclusion

Our results indicate that existing ES benchmarks (e.g., Cohen, 1988) do not depict findings in applied psychology. Specifically, results indicate that the distribution of effect sizes exhibits tertile partitions at values approximately one-half to one-third those intuited by Cohen (1988). In addition, results indicate substantial variability in the distribution of effect sizes across research domains and types of relationships. Indeed, benchmarks for relations involving behavior (e.g., attitude–behavior; intention–behavior) are substantially lower than those not involving behavior (e.g., attitude–attitude; attitude–intention). Our results, and our database which we make available online, offer information that can be used to zoom in on the broader types of relationships to obtain finer grained correlational effect sizes at a desired level of generality.

Taken together, these results are useful for producing better informed non-nil hypotheses and, consequently, will likely facilitate future theoretical advancements. Also, our results offer information that can be used to conduct better informed a priori statistical power analyses, leading to more appropriate sample size determination, which will hopefully help mitigate underpowered

research in the future. Our study offers information that can also be used to understand the relative importance of the effect sizes found in a particular study in relationship to others in the same and other domains. Also, in terms of implications regarding interpretation of results, our study offers useful information about which research domains have advanced more or less, given that larger effect sizes indicate a better understanding of phenomena and ability to predict focal outcomes. Regarding future research, our study offers information about research domains for which the investigation of moderating effects may be more fruitful. Also in terms of future research, our results regarding the distribution of effect sizes across domains offer useful information that is likely to facilitate the implementation of Bayesian analysis. Finally, regarding implications for practice, our study offers information that practitioners can use in terms of evaluating the relative effectiveness of various types of interventions. In sum, we see many useful applications of the effect size benchmarks we obtained and others that can be produced using our database, which we think will benefit applied psychology research and practice.

References

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90,* 94–107. doi:10.1037/0021-9010.90.1.94

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. doi:10.1177/0149206310377113

Aguinis, H., & Harden, E. E. (2009). Sample size rules of thumb: Evaluating three common practices. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences* (pp. 269–288). New York, NY: Routledge.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of *Organizational Research Methods*: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods, 12,* 69–112. doi:10.1177/1094428108322641

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. doi:10.1177/1094428109333339

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

Arthur, W., Bell, S. T., Villado, A. J., & Doverspike, D. (2006). The use of person–organization fit in employment decision making: An assessment of its criterion-related validity. *Journal of Applied Psychology, 91,* 786–801. doi:10.1037/0021-9010.91.4.786

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48,* 587–605. doi:10.1111/j.1744-6570.1995.tb01772.x

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology, 99,* 332–340. doi:10.1037/a0034745

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods, 15,* 17–32. doi:10.1177/1094428110392383

Cascio, W. F., & Aguinis, H. (2008a). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology, 93,* 1062–1081. doi:10.1037/0021-9010.93.5.1062

Cascio, W. F., & Aguinis, H. (2008b). Staffing twenty-first-century organizations. *The Academy of Management Annals, 2,* 133–165. doi:10.1080/19416520802211461

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. doi:10.1037/0021-9010.90.5.928

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65,* 145–153. doi:10.1037/h0045186

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

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. doi:10.5465/AMJ.2007.28165855

Cortina, J. M., & Landis, R. S. (2011). The earth is not round ($p = .00$). *Organizational Research Methods, 14,* 332–349. doi:10.1177/1094428110391542

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. doi:10.1037/0021-9010.79.1.67

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

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. doi:10.1111/j.1744-6570.2012.01243.x

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13,* 668–689. doi:10.1177/1094428110380467

Edwards, J. R., & Christian, M. S. (2014). Using accumulated knowledge to calibrate theoretical propositions. *Organizational Psychology Review, 4,* 279–291.

Ellis, P. D. (2010a). Effect sizes and the interpretation of research results in international business. *Journal of International Business Studies, 41,* 1581–1588. doi:10.1057/jibs.2010.39

Ellis, P. D. (2010b). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511761676

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. doi:10.3758/BRM.41.4.1149

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238. doi:10.1037/0022-3514.50.2.229

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40,* 532–538. doi:10.1037/a0015808

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Orlando, FL: Academic Press.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). San Francisco, CA: Routledge.

Grizzle, J. W., Zablah, A. R., Brown, T. J., Mowen, J. C., & Lee, J. M. (2009). Employee customer orientation in context: How the environment moderates the influence of customer orientation on performance outcomes. *Journal of Applied Psychology, 94,* 1227–1242. doi:10.1037/a0016404

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58,* 78–79. doi:10.1037/0003-066X.58.1.78

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. doi:10.1002/sim.1186

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. doi:10.1111/j.1750-8606.2008.00061.x

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). 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. doi:10.1037/0021-9010.85.6.869

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. doi:10.1037/0033-2909.127.3.376

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152. doi:10.1037/a0028086

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in I–O psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 6,* 252–268. doi:10.1111/iops.12045

Kish, L. (1965). *Survey sampling.* New York, NY: Wiley.

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. doi:10.1177/1094428112457829

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97,* 951–966. doi:10.1037/a0028380

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163. doi:10.1037/1082-989X.9.2.147

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141. doi:10.1207/s15327965pli0102_1

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315. doi:10.1037/0033-2909.97.2.307

Parker, S. K., Williams, H. M., & Turner, N. (2006). Modeling the antecedents of proactive behavior at work. *Journal of Applied Psychology, 91,* 636–652. doi:10.1037/0021-9010.91.3.636

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86–89. doi:10.1111/1467-8721.ep10768746

Rhoades, L., & Eisenberger, R. (2002). Perceived organizational support: A review of the literature. *Journal of Applied Psychology, 87,* 698–714. doi:10.1037/0021-9010.87.4.698

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57,* 221–237. doi:10.1037/h0087427

Roth, P. L., BeVier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330. doi:10.1111/j.1744-6570.2001.tb00094.x

Rudolph, C. W., Wells, C. L., Weller, M. D., & Baltes, B. B. (2009). A meta-analysis of empirical studies of weight-based bias in the workplace. *Journal of Vocational Behavior, 74,* 1–10. doi:10.1016/j.jvb.2008.09.008

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. doi:10.1111/j.1744-6570.2008.00132.x

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83. doi:10.1037/0003-066X.40.1.73

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. doi:10.1037/a0023322

Tracz, S. M., Elmore, P. B., & Pohlmann, J. T. (1992). Correlational meta-analysis: Independent and nonindependent cases. *Educational and Psychological Measurement, 52,* 879–888. doi:10.1177/0013164492052004007

Wallace, J. C., Edwards, B. D., Arnold, T., Frazier, M. L., & Finch, D. M. (2009). Work stressors, role-based performance, and the moderating influence of organizational support. *Journal of Applied Psychology, 94,* 254–262. doi:10.1037/a0013090

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi:10.1037/0003-066X.54.8.594

Yu, K. Y. T. (2009). Affective influences in person–environment fit theory: Exploring the role of affect as both cause and outcome of P–E fit. *Journal of Applied Psychology, 94,* 1210–1226. doi:10.1037/a0016403

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. doi:10.1093/arclin/16.7.653

Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61,* 309–348. doi:10.1111/j.1744-6570.2008.00115.x

Received August 5, 2013
Revision received August 11, 2014
Accepted August 29, 2014 ■