Human Resource Management Review xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management☆,☆☆

Deniz S. Ones [a,*], Chockalingam Viswesvaran [b,**], Frank L. Schmidt [c]

[a] Department of Psychology, University of Minnesota, United States
[b] Department of Psychology, Florida International University, United States
[c] Department of Management and Organizations, University of Iowa, United States

ARTICLE INFO

Available online xxxx

Keywords:
Meta-analysis
Replication
Literature search
Open data
Practitioner research
Publication bias
Interrater reliability

ABSTRACT

This might be the most opportune time for Human Resource Management (HRM) to benefit from psychometric meta-analysis. Explosion of empirical research, often with conflicting results, hide important takeaways that can guide evidence-based applications of HRM. The science of HRM can turn to meta-analyses and meta-analytic thinking as the antidote to the so-called replication crisis afflicting social sciences in general. In this paper, we focus on issues and potential problems that may threaten the veracity and usefulness of contemporary meta-analyses in HRM. We contend that these problems must be correctly tackled for meta-analyses to realize their full potential in advancing HRM science and practice. We address the problems of identification and inclusion of all relevant effect sizes, as well as appropriate corrections for unreliability and range restriction. We offer concrete proposals to enable inclusion of unpublished, practitioner research and data in HRM meta-analyses.

1. Introduction

Human Resource Management (HRM), like other applied sciences, has two lofty ideals: (1) creation of cumulative knowledge based on trustworthy empirical evidence and (2) support of applied practices based on evidentiary science. Empirical studies that address applied questions serve both ideals and can grow the knowledge base of the field. The meta-analytic methods used in HRM were originally invented almost 40 years ago to address the problem of validity generalization (see Schmidt, 2015, for a brief history). Situationally varying validities were creating the impression that psychological tests were only valid for specific purposes, in specific situations, and in specific settings. Schmidt and Hunter (1977) tested the hypothesis of situational specificity by statistically pooling available validation studies, correcting for statistical artifacts such as unreliability in measures and range restriction. The techniques they developed to show validity generalization have over time expanded to quantitative summaries of virtually all HRM effect sizes where more than a few studies have estimated a focal relationship, and are now referred to as psychometric meta-analysis. This is in recognition of the fact that the methods introduced by Schmidt and Hunter correct for

☆ All three authors contributed equally; order of authorship is arbitrary.

☆☆ We thank Stephan Dilchert and Brenton Wiernik for valuable review, insights, and editorial assistance. We thank Brenton Wiernik and Casey Giordano for help with compiling supporting references.

* Correspondence to: D. S. Ones, Department of Psychology, University of Minnesota, Minneapolis, MN, United States.
** Correspondence to: C. Viswesvaran, Department of Psychology, Florida International University, Miami, FL, United States.
E-mail addresses: deniz.s.ones-1@tc.umn.edu (D.S. Ones), vish@fiu.edu (C. Viswesvaran).

http://dx.doi.org/10.1016/j.hrmr.2016.09.011
1053-4822/

biasing effects of statistical artifacts such as unreliability and range restriction, whereas other meta-analytic methods only concern themselves with sampling error. Another noteworthy feature of psychometric meta-analysis is that it is based on a random effects model rather than a fixed effects model (Schmidt, Oh, & Hayes, 2009).

Validity generalization studies first illustrated the sizable, distorting impact statistical artifacts have on individual study results, and led to the application of psychometric meta-analytic techniques to virtually all areas of inquiry in HRM research. Indeed, applications of psychometric meta-analysis have led to an epistemological paradigm shift in HRM. The best peer-reviewed journals routinely publish meta-analyses (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011; Shen et al., 2011). Textbooks rely on meta-analyses for summary conclusions. Meta-analyses are cited at greater rates than primary empirical articles (Patsopoulos, Analatos, & Ioannidis, 2005). Nevertheless, the scientific contributions of meta-analysis should not be judged solely by its popularity among researchers, however. The true influence of meta-analysis stems from its ability to answer important research questions with more robustness than conventional empirical research techniques. While meta-analysis may seem most appropriate in fields where precision in estimates is desired and fields where sampling and measurement errors may obscure the significance of observed effects, scientists in myriad disciplines have come to understand that the techniques of meta-analysis reveal true magnitudes of relationships and allow for more stable empirically-based conclusions. Our field is no exception.

Against this backdrop, the past decade has brought changes to the way that research is conducted and disseminated in our field. These include the ability: to gather data from very large samples and accessing big, existing datasets; to input and analyze data with lightning speed; to identify, search for and gather relevant published research articles en masse within a matter of minutes; and to instantaneously share research with colleagues globally (Spellman, 2015). These are indeed profound changes in the way that research is conducted and disseminated. To be sure, they affect all scientific disciplines, yet forces unique to our applied field may exacerbate their consequences for the science and practice of HRM.

For the scholarly community, these changes have brought about an explosion of published research, open source journals, and unprecedented volume of studies from non-English speaking countries. Samples have shifted from traditional employee, applicant, and student categorizations to include paid panels, remote workers on platforms such as Amazon's Mechanical Turk, and participants recruited through other online channels (e.g., Facebook users). Merits and demerits of these samples are being discussed in the literature (Landers & Behrend, 2015). With increased ease in conducting and disseminating primary studies, two unintended consequences have afflicted HRM: greater premium on novelty rather than replicability in published research (Campbell & Wilmot, in press) and theory worship (i.e., placing a premium on theory development and refinement over empirical discovery; Hambrick, 2007).

For practitioners, the volume and velocity associated with big data is transforming how HRM is practiced in organizations. Businesses and consultancies that place a premium on evidence-based HRM practices are gathering and analyzing their own data, and are almost exclusively consuming their own research (e.g., Google; Dekas, Bauer, Welle, Kurkoski, & Sullivan, 2013). Empirical studies arising from practice settings (often conducted by practitioners, not academics) are considered to contain unique information and knowledge, constituting a major source of competitive advantage. It is little surprise then that most discoveries, inventions, and knowledge from the field are increasingly argued to be proprietary.

The wedge between science and practice in HRM has never been greater. In this paper, we argue that psychometric meta-analysis, when properly applied, is a cause for optimism. We first note the proliferation of studies in HRM and note the deleterious effect questionable research practices can have on HRM research. We then highlight the role that psychometric meta-analysis can, and must, play for building cumulative knowledge in our field, as well as the importance of approaching new developments with meta-analytic thinking. In the second half of this paper, we take up two issues that contemporary meta-analyses should carefully attend to, in order to ensure that they live up to their full potential: research inclusiveness and appropriate corrections for measurement error and range restriction. Psychometric meta-analysis is a constantly evolving research integration tool (Schmidt, 2008). Our goal is to provide guidance to future, new meta-analyses and updates to existing meta-analyses in HRM.

1.1. Proliferation of primary studies in HRM

Knowledge about and the prediction of work behavior depends on accurate assessment of the relationships between variables. With the ease of data collection, analysis, and sharing, our field is awash with empirical research. For example, using the search term “job satisfaction” yields 13,992 hits on Social Sciences Citation Index. The terms “leadership”, “personality and performance”, “teams”, “employment training”, and “goal setting” yield 42,357; 10,374; 51,627; 4023; and 2704 articles, respectively (search conducted January 26, 2016). With such a large body of studies, literature reviews of hypothesized relationships require a systematic, objective, and empirical approach to integrate research findings. Without meta-analyses, it is fast becoming impossible to draw generalizable conclusions from these and most other HRM literatures. Meta-analyses are essential as an objective summary tool.

As can be expected, with such voluminous research, conflicting findings are an inevitable consequence of sampling error and other statistical artifacts. That is, when multiple studies are reported, we find that there are conflicting findings for relationships under examination. Some studies report a statistically significant relationship whereas others report null findings—or even statistically significant findings in the opposite direction. We will not belabor the problems associated with statistical significance tests here (see Cohen, 1990, 1994; Guttman, 1985; Harlow, Mulaik, & Steiger, 1997; Kaplan, Bradley, Luchman, & Haynes, 2009; Kline, 2013; Lykken, 1968; Meehl, 1990; Morrison & Henkel, 2007; Nickerson, 2000; Rogers, 2010; Schmidt, 1996; Schwab & Starbuck, 2009; Ziliak & McCloskey, 2007, for excellent treatises) or the structural and institutional changes necessary to eradicate their misleading influence (Orlitzky, 2012), though we note that journal editors have started to institute policies that discourage, if not

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

outright ban reliance on statistical significance (Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016). With and without significance tests, wildly varying effect sizes from primary studies have sent and increasingly continue to send HRM researchers in search of moderators, which in turn are assessed using inappropriate techniques (Schmidt & Hunter, 2014). Even after a voluminous and exhaustive search of moderators, conflicting results abound, given second order sampling error or capitalization on chance inherent in some of the methodologies utilized to detect such moderators.

1.2. Replication and meta-analytic thinking: changing mindsets for an epistemological revolution

Despite the historically unparalleled abundance of empirical investigations, a new concern has been sweeping social science research, often referred to as "the replication crisis" (cf. Stanley & Spence, 2014). The concern is over the supposed failure of social scientists to conduct replication studies, and the failure to replicate initial results in those instances where such studies are conducted. Some researchers—and even some methodologists—see replication as the solution to the current crisis of confidence about the soundness and trustworthiness of our academic literature (cf. Bartlett, 2012; Kepes & McDaniel, 2013, p. 261; Makel, Plucker, & Hegarty, 2012; Novotney, 2014; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012; Ritchie, Wiseman, & French, 2012; Roediger, 2012). They see replication as "the gold standard of science" that much of our literature is lacking.

However, most commentators advocating replication as the gold standard of science typically assume that replications should be interpreted in a stand-alone manner. This is faulty logic. Let us use an example to illustrate. If an initial study finding a significant result is followed up by an attempt at replication that fails to obtain a statistically significant result, this "failure to replicate" is assumed to constitute important scientific evidence casting doubt on the initial significant finding (Schmidt & Oh, 2016). In this way of thinking, failure to replicate is defined as failure to get a significant result, and successful replication is defined as getting a significant result. Such thinking is erroneous because it completely ignores statistical power. Average statistical power in most psychological literatures ranges from 0.40 to 0.50. (cf. Bakker, van Dijk, & Wicherts, 2012; Cohen, 1962, 1977, 1988, 1992; Schmidt, Hunter, & Urry, 1976), a phenomenon that still afflicts more than half the studies in even the best HRM journals (Shen et al., 2011). Such chronically low levels of power mean that a failure to replicate tells us next to nothing. If confidence intervals (CIs) were used instead of significance tests, there would be far fewer "failures to replicate"—because CIs would often overlap, indicating no conflict between multiple studies (Cumming, 2012; Schmidt, 1992, 1996).

Even 40 years after the invention of psychometric meta-analytic methods, the problem afflicting HRM research is a false belief in the law of small numbers and a severe underestimation of the effects of sampling error. Most HRM researchers believe that a random sample from the population of interest is representative of the population from which it is drawn (e.g., job applicants, employees), even if it is a small sample. But in fact, such a sample is representative *only* if it is a large sample. Small samples are randomly unrepresentative (Tversky & Kahneman, 1971). The consequence is that even if significance tests were not used, and confidence intervals were used instead, they would be very wide, indicating each study contains little information.

We believe that replications are helpful because they provide the studies needed for systematic, large-scale investigations of relationships of interest in the form of meta-analyses. Meta-analysis of multiple studies solves the problem of low statistical power and precision in individual studies. The epistemological basis of our field needs to catch up with meta-analytic thinking.

The model of cumulative scientific knowledge that many researchers appear to follow today is defective. It is a sequential model that can be roughly characterized by the following sequence: First, a researcher conducts a study that gets significant results for the initial hypothesis. Then the researcher concludes, "This question has been answered. Now we move on to test a different hypothesis from the theory." The researcher then moves on to test to another hypothesis. If the result is non-significant, the researcher concludes that that hypothesis is disconfirmed and moves on to another hypothesis. If the result is significant, the researcher concludes that the hypothesis is confirmed and moves on to another hypothesis. Following this sequence is often viewed as constituting a "research program". This model makes the false assumption that a single study can answer a question or establish or disconfirm a hypothesis. The goal of scientific inquiry is to establish principles that generalize for a population larger than the original sample under investigation. This can only be achieved via a meta-analysis of multiple studies. Because researchers don't realize how large sampling errors can be, they don't appreciate how much capitalizing on sampling error can inflate individual research findings. Other distorting factors beyond sampling error include measurement error, range variation, imperfect construct validity of measures, and others (cf. Schmidt & Hunter, 2014)—errors that are often poorly understood (or addressed in individual research).

Primary studies and replications need not have statistically significant results to make a contribution. Even if *none* of the studies cumulated in a meta-analysis reported statistically significant results, meta-analysis could still demonstrate the existence of an important relationship. Non-significance in underpowered studies has no informational value. No single study can answer any question, because too many artifactual factors perturb its results. Meta-analysis offers the only solution to creating a reliable body of cumulative knowledge.

We propose that graduate training in HRM methodology emphasize the correct conceptualization that replication plays, and ought to play in HRM research and meta-analytic thinking. Educational efforts should also be directed to researchers and practitioners who are past their doctoral training to help reshape mindsets for the epistemological revolution necessary to rely exclusively on meta-analyses for scientific discovery and theory building. Journal editors, reviewers and professional workshop/training organizers have serious responsibility to communicate, educate and promote meta-analytic thinking. Publication of articles such as this one can also help in this enterprise. Accordingly, in the next section, we lay out the correct role for individual primary studies and the role of meta-analysis for HRM science.

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

1.3. Meta-analysis as a powerful solution

No individual study, however well designed, can definitively answer any question; thus there should be no attempt to draw conclusions from any such individual study in isolation. Multiple studies should be conducted on each relation, hypothesis, or effect, using meta-analysis as a solution to the problems of sampling error and low power and precision in individual studies (Braver, Thoemmes, & Rosenthal, 2014; Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 2014). The epistemological model in HRM research should be based on meta-analyses because the scientific meaning of a corpus of studies is revealed in the results of such meta-analyses.

The good news is that, as discussed above, empirical research in HRM is more voluminous than ever. The velocity of production of primary studies is faster than ever. And the large number of meta-analyses in our literature shows that replication studies are, in fact, being conducted in most areas of HRM research. Indeed, as the years pass, several seminal meta-analyses in our field can be updated by integrating new primary studies that have been conducted since the publication of the respective originals. However, all these developments place an onus on meta-analyses and the quality with which they are conducted.

Meta-analysis can only serve the function that it needs to serve—providing solutions to the problems identified above—only if individual meta-analyses reflect the entire research literatures, without bias. The real problem is not unavailability of studies or a lack of replication of research; it is the potential distortion of our research literatures reflected in meta-analyses. In other words, meta-analyses first must have available a full, undistorted range of primary studies to use as their input. Second, meta-analytic methods should correct for the appropriate statistical artifacts using appropriate techniques, or risk producing biased results.

On these points, we raise two issues that reduce the potential of meta-analyses to inform our science and practice. We first discuss the issue of inclusiveness in meta-analytic cumulations, enabling meta-analyses to have full, undistorted literatures to synthesize. Beyond publication bias and questionable research practices, we point out that many recent meta-analyses in HRM fail to include full sets of studies that exist on a topic, resulting in a distribution of observed effects where systematic swaths of literature may be lacking—skewing results beyond any potential publication biases that have been discussed in the recent literature. Second, we discuss the necessity to use the correct unreliability and range restriction corrections to estimate effects without systematic biases. [The issue of using appropriate moderator search approaches is also crucial, but is addressed elsewhere (Schmidt, in press), and is therefore outside the scope of this manuscript.]

2. Undistorted, full range of primary studies for unbiased meta-analytic databases: publication bias and questionable research practices

Leaving aside outright research fraud (Fountain, 2014; Matlack, 2013; Schmidt & Hunter, 2014, pp. 521–523; Verfaellie & McGwin, 2011), publication bias is an important issue in many research literatures that can distort meta-analytic results. If publication bias evidence is found, sensitivity analyses should be reported (e.g., see Schmidt & Hunter, 2014, Chapter 13). It is not the purpose of this paper to review the large literature on publication bias that has emerged (e.g., Rothstein & Hopewell, 2009; Rothstein, Sutton, & Borenstein, 2005) or its effects on the HRM literature (e.g., Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012); interested readers may consult these primary sources.

In addition to publication bias, meta-analyses may be biased due to effects of questionable research practices in primary studies (see John, Loewenstein, & Prelec, 2012, Simmons, Nelson, & Simonsohn, 2011, for recent surveys; cf. Schmidt & Hunter, 2014, pp. 523–530, Chapter 13). Examples include: (a) adding subjects one by one until the result is significant, then stopping; (b) dropping studies or measures that are not significant; (c) conducting multiple significance tests on a relation and reporting only those that show significance; (d) deciding whether to include data after looking to see the effect on the significance outcomes; (e) harking: hypothesizing after the results are known; (f) running an experiment over again until the "right" results are obtained. John and colleagues found that large percentages of the researchers they surveyed admitted to using these practices.

Severe problems of questionable research practices have been detailed elsewhere (e.g., Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995; cf. Schmidt & Hunter, 2014, pp. 523–525, Chapter 13). We wish to highlight one area of concern that sometimes is overshadowed by more sensational, questionable practices in discussions: Small randomized sample sizes do not produce equivalent groups (Tversky & Kahneman, 1971), and therefore cannot provide a robust basis for inferring causality.

Questionable research practices and publication bias distort the results of meta-analyses (e.g., Bakker et al., 2012; Levine, Asada, & Carpenter, 2009). The typical effect is an upward bias in mean effect sizes and correlations. Combatting this is the task of the entire field (including journal editors and reviewers) and not just meta-analysts. *We propose that (a) journal editors and reviewers should be vigilant against questionable research practices, (b) meta-analysts should be vigilant against publication bias, and (c) our field should value strict and conceptual replications and multiple studies, big and small, addressing the same or similar questions as a safeguard against distorted meta-analytic findings.*

2.1. Reflecting the scholarly literature in meta-analyses

Now we turn to a discussion of unrepresentativeness of some meta-analyses with regard to primary studies they include. To assess the relationship between any two variables, meta-analysts need to locate and obtain a comprehensive set of studies documenting the relationship. We suspect that most of the published HRM meta-analyses are not comprehensive in their coverage. Let us give two recent examples. First, Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012) aimed to update an earlier meta-analysis on the validity of integrity tests (Ones, Viswesvaran, & Schmidt, 1993) by including earlier research and newer

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

studies. Yet, their meta-analyses included just over half the studies included in the previous meta-analysis. Only one third of the studies they included overlapped with the earlier meta-analysis (Ones, Viswesvaran, & Schmidt, 2012). A second example comes from the two recent meta-analyses of the interest literature (Nye, Su, Rounds, & Drasgow, 2012; Van Iddekinge, Roth, Putka, & Lanivich, 2011). Van Iddekinge et al. comprehensively examined the relations of job and training performance and turnover with all forms of interest measures as they are used in practice (including construct-based, basic, and occupational interest scales, applied as either single job-relevant scores, interest-occupation congruence indices, or regression-based composites), whereas Nye et al. limited their search and analyses to a single theoretical perspective and a limited set of analytic methods used by researchers and practitioners (i.e., RIASEC construct scales, applied as either single scores or congruence indices). As a result of their broader scope, Van Iddekinge et al. included more than twice as many studies as Nye et al. and were able to much more comprehensively address the state of knowledge on vocational interests and work performance. In most cases, Van Iddekinge et al. and Nye et al. came to diametrically opposed conclusions. Van Iddekinge et al. found small to moderate validity for single job-relevant interest scales and congruence indices and moderate to large validity for regression-based composites; Nye et al. found large validity for single (even job-irrelevant) interest scales and even larger validity for interest congruence indices. Nye et al. presented their meta-analytic methods without sufficient clarity to fully determine the source of the discrepant results (though we expect the uniformly larger validities may be an artifact of their use of meta-regression, rather than traditional meta-analytic methods, to conduct their analyses).

What are the likely pitfalls in identifying, obtaining, and utilizing all available data for a meta-analysis? First, many researchers have become indolent, relying only on electronic searches for studies and data sources. Depending on the electronic databases chosen for the search, there is potential to miss relevant studies. For example, reliance on only one or a few databases such as PsycInfo or ABIinform likely results in missing studies from uncovered domains such as medicine (e.g., HRM research on medical doctors published in medical journals; HRM research on nurses published in nursing journals, HRM research on police officers published in criminology journals, HRM research on service employees published in hospitality journals and so forth). HRM meta-analyses may underrepresent primary studies on some occupational groups unless a conscious effort is made to include them. On a related note, all databases do not cover the same time period and as a result may miss additional sources. There is also a concern that much of the research dating before 1960 is missing from HRM database searches. We note that earlier studies, especially field-specific ones, were often published not in journal articles, but summarized in books (e.g., Strong, 1943; Tiffin, 1948) or published as entire books dedicated to one large study (e.g., Burt, 1917). Even in recent literature, we can think of some very valuable research being published in the form of a book rather than several individual journal articles. For example, Project A was a major HRM study undertaken in the U.S. Army. Some early results were published in *Personnel Psychology* (e.g., Campbell, 1990) and the *Journal of Applied Psychology* (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). However, the project was too large and produced too many findings to be published piecemeal in our journals. Campbell and Knapp (2001) assembled some of the key studies in an edited volume. Other areas of our discipline have produced similar efforts (for example, research on creativity and innovation during the 1950s and 1960s). If researchers exclusively rely on electronic searches of conveniently available databases, much primary research will be missed. For example, not including Project A studies can result in systematic absence of valuable military research from a meta-analysis. Deep knowledge of past and ongoing research in a content area serves the meta-analyst well, as non-journal study repositories can be evident to the expert, but not the novice. *We propose that meta-analysts do not limit themselves to electronic searches or searches of only contemporary journals. When electronic searches are undertaken, we recommend the use of multiple databases that span multiple fields, and not just HRM or applied psychology. Electronic searches for books containing relevant data should be supplemented by manual searches of pre- 1960s I-O psychology and HRM related books as well as books identified by field experts as relevant.*

Second, we cannot underemphasize the importance of snowballing (i.e., using reference lists of coded studies to identify additional primary studies). In this enterprise, combing through the citation lists of important reviews and highly cited articles is also fruitful. *We propose exhaustive snowballing as essential to all meta-analytic efforts.*

Third, identifying relevant dissertations and conference papers has become relatively easy thanks to electronic availability of conference programs and dissertation databases. However, obtaining them can be an effortful enterprise. Dissertations may or may not be available electronically. Those that are available may cost a non-negligible sum to purchase. Conferences papers are sporadically available, and are often abstracted without quantitative results. Diligence in contacting authors of conference papers and dissertations is an essential strategy for the meta-analyst. *We propose that meta-analytic efforts should obtain and include conference papers and dissertations in their databases, unless there is a compelling reason for not including unpublished conference papers and dissertations among studies to be pooled* (e.g., *when the authors are aiming to examine only published findings*).

Fourth, despite globalization of business and human resources practices, HRM meta-analyses typically restrict themselves to primary studies reported in English language from developed countries. To be sure, some meta-analyses include data from different countries and report international comparisons (e.g., Yoon, Schmidt, & Ilies, 2002). However, the search and location of primary studies from electronic databases remains an English-only enterprise. Even studies written in English from Asia, Middle East, and Africa are poorly indexed in large databases. Despite increasing appearance of non-Anglo-American research in our journals (Shen et al., 2011), most meta-analysts do not include research from non-English sources. *We propose that meta-analysts should (1) search databases in other languages, (2) work with translators, native speaker research assistants and collaborators, and at the very least (3) use translation software* (e.g., *Google Translate*) *to access primary research in different languages.* Our students and we have had great success with these approaches uncovering caches of studies not included in previous meta-analyses we were updating (Stanek, 2014; Wiernik, 2016). The science and practice of HR is a global enterprise and we should integrate research from different languages. Meta-analyses that include data from different countries and languages are especially

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

valuable in testing cross-cultural generalizability of relations and effects, opening the door for a knowledge base that is sensitive to universal fundamentals vis-à-vis potential local/country/language/culture specific theories and conceptualizations (see Ones, Dilchert, Deller et al., 2012, for detailed explanations and examples of meta-analytic cross-cultural generalization research).

2.2. Reflecting practitioner data in meta-analyses

Human resource management is an applied science. Evidence based practice requires empirical research that uses tools and practices as they are employed in the field. As an applied science, our most valuable data are field data. Practitioners that work for specific companies in industry or governmental bodies, for consulting companies that offer services to organizations, and for nonprofits, have access to such precious field data. Many conduct primary studies of their own to assist in organizational decision making (e.g., Which selection tools to use? Is a given intervention acceptable to participants? Is it effective?). It is an unfortunate reality that unless practitioners place their research in the scientific literature by publishing or presenting it, these data are entirely missing from most meta-analyses conducted in HRM. As a consequence, meta-analyses can be systematically biased because they do not include much, if not most, of relevant field data. This is certainly true in areas such as training effectiveness. It is standard practice to assess at least participant satisfaction in training interventions. Yet, we do not have a vast field-based literature on trainee reactions to training. Many educational programs assess learning in some form or another. Yet, our training effectiveness literature is based on no more than a few hundred studies. Similar points can be made with regard to other HRM interventions, including employee selection procedures (e.g., why do meta-analyses of employment interviews include only a few dozen studies, when virtually every organization conducts interviews pre-hire?).

There have been two impediments to industry practitioners sharing their data with meta-analysts. First, practitioners do not have the time or the incentive systems to prepare and share databases. This can be overcome by academicians, graduate student interns, or meta-analysts partnering with practitioners. Second, organizations are protective of their data. Shared data can reveal organizational vulnerabilities and create potential legal liabilities. Shared data can also give away proprietary knowledge which constitutes competitive advantage. Yet, our experience has been that when organizations *trust* the meta-analysts, when they can be assured that their data will be appropriately protected, productive data sharing arrangements are the result. Many are willing and some even proactive about sharing their data with academics who they believe have the competence to correctly analyze data and whom they trust to be unbiased. Meta-analysts must earn the trust of practitioners working in industry while upholding professional standards of research conduct.

But practitioners must also be convinced of the value of meta-analytic research for their organizations, and the contribution of shared data to the greater good, including economy-wide productivity and employee well-being. As organizational data increase in size, practitioners contend that their need for meta-analyses from scholars diminishes. Yet, this is short-sighted. All organizational data, even of gargantuan proportions, reflect only the here and now. Findings from such data are bound by the specific measures utilized as well as the current conditions and workforce composition in the organization (for example, depending on geographic location or complexity of jobs included, results might be applicable only to certain demographic groups). Meta-analyses, by pooling data from multiple organizations, settings, and conditions, reveal (a) generalizability of findings, and—more important in this case—(b) reveal key features that influence findings. Testing limits of theories, laws, and corollaries is a very important part of science. The only way to determine how the effectiveness of HRM interventions changes (or stays the same) is by pooling data from as vast conditions and settings as possible and sorting out potentially moderating influences meta-analytically. Using only single organizational data, however large, to predict future usefulness of HRM interventions, given different organizational, occupational, and industry conditions, is a fool's errand.

Consulting companies and test publishers also have vast caches of data from applied settings, often documenting the nomological net of various products and services they market. This is as it should be and bodes well for evidence-based HRM. For example, several consulting companies have developed predictors for personnel selection and training programs for various interventions. It is essential for meta-analyses to include findings from consulting and test publisher studies and databases. In fact, they may be more important than published studies in building cumulative knowledge for our *applied* science. For example, public domain measures (e.g., IPIP scales) examined in lab or traditional research settings (e.g., with college student or MTurk samples performing simulated tasks in context-free settings) may not have the same nomological network as their carefully constructed counterparts designed to meet organizational demands (e.g., faking resistance for personality measures, appropriate reading levels, items screened for DIF for various groups, work-contextualized items, and so forth).

An extreme illustration is a scenario where no one but consulting companies may have the relevant data. For example, with the trend of gamification in employee selection, high fidelity simulations are becoming increasingly common. Such simulations employ 2D and 3D animations, avatars, and so forth. Yet they are extremely expensive to develop (Fetzer & Tuzinski, 2013), and analogues in research settings either lack realism or are nonexistent. The only way to learn about (1) gamification features and their psychometric characteristics, and (2) the extent that validities of such simulations generalize across occupational groups and various settings is through cumulating consulting company data in a comprehensive meta-analysis. In today's peer reviewed publication zeitgeist, empirical investigations of the constructs measured, their criterion-related validity, or utility of the tools used to measure them, are unlikely to appear in our top journals; they are often deemed to contribute little to our science and theory. Often, data sit in consultant files. At best, technical reports may be available. Thus, if these are not included in meta-analyses, we are sure to get a distorted view of important relationships from meta-analyses.

We propose that the default baseline for HRM meta-analyses should be to accept and include all practitioner data in HRM meta-analyses, without any preconceived notions. Arguments that consultant and vendor data untrustworthy is without merit and should

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

be rejected. Professional standards even *require* data from test publishers to document empirically supported use of psychological tests (AERA, APA, & NCME, 2014).

2.2.1. On the acceptability of consultant and product vendor research

Although recent discussions of academic research, published in the peer reviewed literature has pointed to troubling biases and distortions of conclusions due to questionable research practices and major shortcomings due to overreliance on statistical significance testing (see above), academics in our field have developed, in our view, an unjustified, disastrous mistrust of consulting and vendor data that is affecting the quality of meta-analyses (cf. Ones, Viswesvaran and Schmidt, 2012). This mistrust arises from claims of biased research, poor quality research, and suppression of null findings. We address each of these in turn.

2.2.1.1. Potential for biased research. A major claim is that consultant research may be tainted by a profit motive. The analogy is with the medical domain where research funded by pharmaceutical companies has been criticized as biased regarding drug effectiveness. Yet, the medical analogy is flawed for the following reasons. Drugs are tested for specific ailments. However, HRM products of consulting companies and test vendors are developed and used for a variety of applications. For example, 360-degree rating tools are used for development, feedback, coaching, promotions, and even job assignments. Tests are developed to predict multiple criteria, for example, task performance, avoidance of counterproductive behaviors, or improving overall work performance. Similarly, tests might be employed for multiple purposes, such as counseling, development, selection, and etcetera. Validation is undertaken for many different criteria.

References to "the validity" of a test are professionally outdated by about 30 years and are simply wrong. Tests and HRM commercial products are expected to be supported by many different lines of evidence that explicate their construct validity and usefulness in the workplace. We are not aware of a single test publisher or vendor that presents only evidence for a sole specific criterion; many lines of evidence are typically reported. As such, in product validation efforts, multiple criteria of varying specificity (e.g., interpersonal facilitation, organizational citizenship, engagement, task performance, absenteeism, general counterproductivity) and domain (e.g., overall job performance, turnover) are typically included. Relationships with other similar and dissimilar measures are presented. Group differences are noted. This means that across test publishers and criteria, there is no single effect that can be the target of manipulation and validity maximization efforts. The plurality of product providers and the plethora of purposes for which measures, tests, and interventions are designed and used discount the possibility of systematic misreporting. Even if a given consultancy were to produce upwardly biased validity estimates for a given criterion, this bias would vary across vendors and across criteria. Unless there were collusion across dozens of consultancies to report only favorable validities for a specific, previously agreed upon criterion, systemic bias in consultant studies is extremely unlikely. It would be akin to all researchers within HRM to agree to suppress null results on only one specific predictor-criterion relationship, or group difference and so forth. Moreover, professional and legal guidelines within our field, which require practitioners to support their instruments with empirical evidence, are being reduced to absurdity by those whose attitude it is to later categorically reject any such evidence in principle. Practitioner data, research, and evidence are necessary and desirable, and comprehensive meta-analyses reflecting the full plurality of available data best estimate the operational effectiveness of commercial HRM products.

Though previously hardly discussed, our field is also becoming increasingly aware that a potential profit motive is not in the least limited to the applied community of our field. Not only do academics have their favorite constructs (and sometimes their own proprietary measures and undisclosed conflicts of interests), but their career succession and successes (and therefore, livelihood) are often tied to the publication of significant findings, as well as evidence consistent with their (or current mainstream) theories (Campbell & Wilmot, in press). Everyone has a stake in the results of their research *and* their applied work. The solution is not to discount one type of research a priori, but to (1) build incentives for everyone to follow standards of applied and scholarly research more, and (2) trust that diversity of interests and research will be the invisible hand providing fuel to meta-analyses to reveal knowledge.

Finally, it may be true that research conducted to address some organizational need tends to yield higher effect sizes than studies conducted for research purposes alone. This is likely because interventions work best when they are needed (e.g., productivity gains are greatest when productivity levels are low; predictors of counterproductivity yield validity when there is counterproductivity to be predicted). Furthermore, we would expect that test creator/consultant expertise is likely to result in better HR products and services compared to those created by novices or for use in single research studies. Job knowledge and experience are time-proven predictors of work performance (Schmidt, Hunter, & Outerbridge, 1986); consulting and test development are not exceptions. Stronger effects reported by consulting companies can also be attributed to product creation and implementation expertise in solving real-world HRM problems.

2.2.1.2. Potential for suppression of studies. The meta-analysts' fear here is the suppression of entire studies of less favorable research. That is, consultancies and product developers may suppress in-house studies revealing weaknesses in their products and services. While it should be clear from the above that such concern shouldn't be unique to vendor research, it nonetheless merits discussion.

McDaniel, Rothstein and Whetzel (2006) examined publication bias in publisher research using funnel plots and trim-and-fill procedures. On average, across 4 test vendors and 18 scales, the mean difference between the criterion-related validity obtained from meta-analyzing test publisher data and the validity obtained from the trim-and-fill publication bias analysis procedure was 0.03 ($SD = 0.04$). Whetzel (2006) examined and found no publication bias in customer service scale validities. Similarly, no publication bias has been found in investigations of the Big Five personality dimensions (McDaniel, Hurtz, & Donovan, 2006) and the

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

General Aptitude Test Battery (Vevea, Clements, & Hedges, 1993). In a more recent examination of the same issue for a specific integrity test, Pollack and McDaniel (2008) found that the reported, observed validity mean across 70 coefficients was 0.25, whereas the distribution suggested by the trim and fill procedure resulted in a mean observed correlation of 0.23. These negligible differences do not support any systematic suppression of results by these parties. In other domains, when meta-analysts have asked questions that could reveal unfavorable facts about their products, test publisher research has not been any more favorable than the rest of the literature. For example, when race differences on cognitive ability tests were meta-analytically investigated, Roth, Bevier, Bobko, Switzer, and Tyler (2001) found comparable racial differences for a particular ability test reported by the test publisher compared to the rest of the studies from the literature. Across the board, there do not appear to be notable effects inflating validities or masking unfavorable results from test vendor research due to suppression of findings.

2.2.1.3. Potential for poor quality research. It has also been argued that consulting practitioners and product vendors make poor methodological choices, leading to distortions in research findings. Arguments have been made that quality studies alone should comprise the meta-analytic database. For example, Morgeson et al. (2007) have stated "gathering all of these low quality unpublished articles and conducting a meta-analysis does not erase their limitations. We have simply summarized a lot of low quality studies" (p.707). The argument is not unique to meta-analyses, traditional narrative reviewers have usually narrowed their reviews to a few handful of studies with arguments that the other studies were methodologically flawed, as well. In addition, we might add that meta-analytic techniques actually *do* remove methodological limitations from poor quality studies – sampling error, measurement error, and artificial dichotomization or variables chief among them.

All data have a place in scientific inquiry. For example, much maligned self-reports of behavior provide construct validity evidence for psychological measures used in employee selection. If certain measurement, study design, or data collection features are expected to produce biased results, these can be identified and quantified in meta-analyses. Inclusion of all available research in meta-analyses can enable moderator analyses examining studies which do not meet some pre-specified quality rule (e.g., random assignment of subjects), often resulting in more informative conclusions. Meta-analyses can use all available evidence, carefully separate and analyze identifiable features indicating quality, and inform both science and practice of HRM accordingly.

For over a century we have been a field led by applications—theories must serve the solution of applied problems, not be a goal by themselves (Campbell & Wilmot, in press). Practitioners were literally serving in the trenches during World War I, when their testing innovations created the science of HRM. Practitioners today continue to be in the trenches addressing the most important assessment, testing, and HRM needs of organizations on a global scale. Academicians must partner with them to conduct relevant and useful research. We are an applied field. Without field data (which practitioners are often best suited to gather), there would be no science of HRM. This is why any meta-analysis purporting to assess the usefulness of any HRM intervention or test must comprehensively include practitioner research and data.

2.2.2. Accessing consultant and product vendor data for meta-analyses

Large proportions of test validity databases are not in professional journals because most of our journals are not in the business of publishing validation studies. Several decades ago, *Personnel Psychology* abolished its Validity Information Exchange section, and journals and reviewers increasingly moved to requiring novel theoretical contributions from submitted studies. With the exceptions of the *International Journal of Selection & Assessment's* research report section, *Educational & Psychological Measurement*, *Applied HRM Research*, and the recently launched *Personnel Assessment and Decisions*, the reality of applied psychology and HRM journals today has been that simple validation studies and replications demonstrating validity of personnel selection tests, however well-conducted, have not been typically publishable. New theoretical contributions or new empirical discoveries are essential for publishing primary studies in professional journals. "What is new here?" or "What is the theoretical contribution?" are questions frequently encountered in the journal review process and often the cause of rejections. Very few practitioner studies are submitted for publication at leading journals and those that are submitted tend to be rejected *not* due to their poor quality, but rather because of our journals' kowtow to the primacy of theoretical contributions and an unfortunate diminishing regard for science through multiple replications that can feed objective data synthesis. Practitioners produce solid validation research, just not in the peer reviewed literature. Rather, publically available technical manuals, in house technical reports/summaries, and data must be accessed by meta-analysts to ensure inclusion of the full range of existing studies for the purpose of meta-analysis.

We must note, however, a few potential obstacles that must be overcome. The technical reports from vendors, being written for documentation, are unlikely to have all necessary information, and thus the meta-analyst needs to rely on the vendor for taking the time to answer potential queries. In the competitive world of consulting, this is likely to place some burden on the vendor. To compound this, issues of client confidentiality, other potential legal issues, and proprietary rights must be contended with (cf. Harris et al., 2012).

All in all, a real threat to meta-analyses, and therefore to cumulative knowledge, is the tendency of most meta-analyses in our field not to include all available empirical evidence on a question, published, unpublished, from academicians, practitioners, consultants, and so forth. This is a major problem that needs to be addressed. Problems around data inclusion and reporting in meta-analyses must be addressed because they threaten the accuracy of the results of meta-analyses, and meta-analyses are the only real foundation for cumulative knowledge.

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

3. Some proposals for collection and sharing of practitioner data for meta-analyses

We are supportive of the idea of open data in HRM research. Yet, we acknowledge there may be legitimate exceptions. As is noted in a universal accord on open data, produced by collaboration among the International Council for Science (ICSU), the InterAcademy Partnership (IAP), the World Academy of Sciences (TWAS), and the International Social Science Council (ISSC), "although open data should be the default position for publically funded research data, not all data can or should be made available to all people in all circumstances. There are legitimate exceptions to openness …Exceptions to the default should be made on a case-by-case basis, with the onus on a proponent to demonstrate specific reasons for an exception." (ICSU, ISSC, TWAS, & IAP, 2015, p. 5).

We distinguish between two types of data and information (1) proprietary features of products and services (e.g., scoring keys, decision algorithms, protocols associated with products and interventions) that are commercially patentable or otherwise valuable, and (2) results of research examining their effectiveness and nomological net. The former can justify exceptions to openness, the latter should not. In fact, for psychological tests, the latter are required by professional standards (AERA, APA, NCME, 2014). Meta-analyses that include uncensored practitioner data serve a key function in self-correcting science.

However, simply calling for making data accessible is not sufficient. We would like to extend the concept of "intelligently open" HRM meta-analyses. The accord quoted from above lists five qualities that define intelligently open data. Below, we situate these in the context of HRM meta-analyses aiming to include unpublished practitioner studies and data. We then use principles of intelligently open data to derive concrete proposals for various stakeholders.

Major qualities of intelligently open data are:

1. Discoverability: In order to be included in meta-analyses, consultancy and practitioner studies, technical reports, and data should be discoverable.
2. Accessibility: Consultancy and practitioner studies, technical reports, and data should be accessible by qualified meta-analysts.
3. Intelligibility: Studies, technical reports, and documentation accompanying data files should contain relevant information for accurate coding by the meta-analyst.
4. Assessability: Providers of technical reports, unpublished studies and datasets should be able to freely assess the competence of data requesting meta-analysts or "the extent to which they may have a pecuniary interest in a particular outcome." (ICSU, ISSC, TWAS, & IAP, 2015, p. 4)
5. Usability: Practitioners should enable usability of technical reports, unpublished studies and relevant data by including adequate metadata, and if necessary, by redacting minimal proprietary or confidential information.

To achieve the qualities of intelligently open data for HR meta-analyses, new hard and soft infrastructures are needed. Here we propose approaches that can be grouped into three major stages in the research process: (a) data and study sourcing, (b) archiving and access, and (c) publication. For each stage, we provide proposed actions for various stakeholders, including practitioners, funding sources, meta-analysts, test publishers, and professional organizations, among others. Table 1 enumerates 21 concrete proposals for all three research process stages, specified for a multitude of stakeholders.

These proposals can be used as a starting point for debate and consideration and can thus contribute to the creation of hard and soft infrastructures as well as future standard practices for reflecting practitioner data in HRM meta-analyses.

4. Appropriate psychometric corrections in meta-analyses

Psychometric meta-analyses in our field start with an observed distribution of effect sizes and correct the observed effect sizes for statistical artifacts such as unreliability and range restriction in the measures (among others). If the observed variability in effects sizes is not satisfactorily accounted for by sampling error and variations in other statistical artifacts across studies, a search of moderator variables is justified. Sampling error is typically the chief source of observed variability across studies. Whereas the previous section of this manuscript focused on increasing the sample size in meta-analyses by including *all* available evidence and thereby minimizing sampling error, we now turn our attention to *systematic* biases that must be addressed using the meta-analytic method itself.

In correcting for systematic statistical artifacts (e.g., measurement error, range restriction) in meta-analyses, individual corrections can be made at the primary study level prior to pooling of results across studies (e.g., Hunter & Schmidt, 2004, Chapter 3; Raju, Burke, Normand, & Langlois, 1991), or by applying distributions of statistical artifacts (e.g., range restriction values, reliability coefficients) to the distribution of effect sizes being meta-analyzed (e.g., see Callender & Osburn, 1980; Hunter & Schmidt, 2004, Chapter 4; Raju & Burke, 1983; Schmidt & Hunter, 1977). In either case, corrections for statistical artifacts are necessitated by the need to remove artifactual, typically downward (range enhancement is an exception), biases affecting primary studies.

Although statistical artifacts distorting study findings include measurement unreliability (Schmidt & Hunter, 2014), range restriction/enhancement (Sackett & Yang, 2000), artificial dichotomization of variables (MacCallum, Zhang, Preacher, & Rucker, 2002), and imperfect construct validity (Hunter & Schmidt, 2004), meta-analyses in HRM consider most frequently the influences of measurement unreliability and range restriction. Although artifacts other than sampling error do not typically contribute greatly to observed variability, they result in substantial underestimation of effect sizes.

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

Table 1
Proposals for reflecting practitioner data in HRM meta-analyses.

Stage in research process	Stakeholders	Proposed actions	Potential benefits
Data/study sourcing	Test publishers and consultancies	1. Report validation results including information on validity study templates provided by Association for Test Publishers	Ensures non-onerous, brief reporting of essential results from validation studies without violating client confidentiality.
	Consultancies, companies and organizations	2. Partner with suitable, potentially vetted, researchers for data analysis and report creation	Enables practice-based data to contribute to science and potentially scholarly output of partner researchers.
	Faculty members and graduate students	3. Encourage students writing theses and dissertations to build partnerships with practitioners and use data from consultancies and business organizations, if appropriate	Enables practitioners to partner with academics and brings practitioner data/findings to searchable and relatively more accessible academic outlets (i.e., dissertations and theses).
	Funding sources	4. Provide funding for retrieval and archiving of technical reports, practice-based studies, and databases from industry and practice	Improves accessibility of grey literature.
	Meta-analysts	5. Specify approaches for obtaining and including practitioner data and reports in meta--analytic methods	Provides an explicit verification of meta-analysts' bona fide efforts for including practitioner data.
	Scientific community	6. Accept practice-based data as scientific evidence without preconceived notions	Enables empirically based meta-analyses that reflect practice; narrows science-practice gap.
Archiving and access	Consultancies, companies and organizations	7. Contribute data and reports to legitimate, publically accessible databases, repositories and archives, after redacting proprietary information and individual identifiers	Reduces burden of internal archiving of data and reports, and makes them accessible to the scientific community at future points in time.
	Independent reviewers of psychological tests (e.g., Mental Measurement Yearbook & British Psychological Society)	8. Build archives of technical reports on reviewed tests that can be accessed by researchers, meta-analysts, and consumers of test reviews	Ensures long term archiving and availability of technical reports that may have been shared with objective reviewers of tests to other researchers and practitioners. Improves future accessibility of test publisher research for scientific purposes.
	MetaBus (Bosco, Uggerslev, & Steel, 2016)	9. Build MetaBus (Bosco et al., 2016) system capability to accept intercorrelation matrices and effect size estimates from practitioners	Makes practitioner data available (in summary quantitative form) for on demand meta-analyses. Makes practitioner findings accessible at future points in time.
	Professional organizations (e.g., SIOP, EAWOP, ITC, SHRM, AOM, ATP)	10. Build repositories for housing data and findings from consultancies and companies (with identities of contributing companies redacted, if necessary) and build online tools for accessing such repositories	Makes proprietary data and/or research studies available to the field and meta-analysts.
	Professional Practice-Based Organizations (e.g., ATP, SHRM)	11. Educate and/or credential meta-analysts for accessing practitioner data and reports containing proprietary information (i.e., on meeting meta-analytic standards while treating proprietary/confidential information with sensitivity information; on managing potential conflicts of interest)	Increases competency and transparency of meta-analysts handling sensitive practitioner data and reports.
	Providers of electronic library databases (e.g., PsychInfo, Social Sciences Index)	12. Build accessible electronic databases to house unpublished technical reports and conference papers; enable submission and archiving such materials by authors/consultancies/-organizations (e.g., by creation or use of uniform reporting templates)	Enables long term access to unpublished materials for broad audiences. Uniform reporting templates improve consistency of information from different sources.
Publication	Journal editors	13. Value and publish strict and conceptual replications from practitioners (i.e., do not solely emphasize empirical studies testing novel theory)	Brings empirical data on actual HR applications and practices to build scientific understanding, including data based (inductive) theory building.
	Journal editors	14. Value and publish null findings	Provides an uncensored distribution of relationships, effect sizes for meta-analytic cumulation.
	Journal editors	15. Value and publish real world problem-driven research	Brings empirical research on and from the field to the literature.
	Meta-analysts	16. Report potential conflicts of interest and indicate sources of support (financial or otherwise) for the meta-analysis	Identifies potential problems with meta-analysts' impartiality.
	Meta-analysts	17. Make meta-analytic data available to interested researchers for verification	Facilitates verification of meta-analytic results, and perhaps more importantly, makes summaries of practitioner findings available to the field.
	Meta-analysts	18. Specify all grey literature sources (e.g., technical reports, company data, author contacts) included in a meta-analysis with	Ensures that grey literature has been sought and specifies how such literature has been obtained; provides others with the means to cull the same

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

Table 1 *(continued)*

Stage in research process	Stakeholders	Proposed actions	Potential benefits
		dates of coverage as well as source/contact information	information from original sources.
	Meta-analysts	19. Indicate processes used for obtaining raw data files (if any) from consultancies and organizations as well as for confirming details associated with each given database (e.g., occupational group sampled, predictive/concurrent validation strategy)	Adds a degree of transparency to the raw datasets used to generate effect sizes included in meta-analyses.
	Meta-analysts	20. Describe approaches used for minimizing systematic bias in grey literature included (e.g., censoring of nonsignificant findings); specify whether bias assessment will be done for each research source (e.g., test publisher) or for each criteria/-outcome (e.g., absenteeism, job performance)	Helps explicitly address any systematic bias (selective reporting of results) in meta-analyses; makes source of any potential bias explicit.
	Meta-analysts	21. Consider including individuals of opposing viewpoints on meta-analysis publication teams to guard against potential biases	Provides healthy skepticism for all meta-analytic decisions.

4.1. Appropriate unreliability corrections in meta-analyses

Measurement error in scales assessing study variables systematically biases relationships downward. The relationships between variables is reduced in size to the extent that both variables are assessed with less than perfect reliability (Lord & Novick, 1968; Schmidt & Hunter, 1996, 1999). Unreliability in the measures is inevitable and corrections are needed (cf. Schmidt & Hunter, 1996; Schmidt, Le, & Ilies, 2003). The correct estimate of reliability depends on the measures used and the substantive research questions asked, especially generalization hypotheses being tested.

When stable individual differences constitute the variables under investigation, coefficients of equivalence and stability provide the appropriate indices to correct for measurement error. This is because random response, transient, and specific factor errors are simultaneously incorporated into coefficients of equivalence and stability (Le, Schmidt, & Putka, 2009; Schmidt et al., 2003). If decisions are to be based on scores on individual differences measures (e.g., selection), operational validities are computed without corrections for measurement error in the predictor.

When HRM variables are measured using raters or evaluated using judges, the source of measurement error variance arises from disagreements among raters. Raters are utilized in pre-employment assessments such as interviews and assessment centers. Raters are prevalent in virtually all assessments of job performance and organizational citizenship behaviors. If employment decisions are to be based on ratings, corrections for measurement error are inappropriate (i.e., observed scores must be used). However, for research and validation, when raters are used to measure criteria, the appropriate reliabilities to be used in attenuation corrections are interrater reliabilities (King, Hunter, & Schmidt, 1980; Rothstein, 1990; Viswesvaran, Ones, & Schmidt, 1996). The goal is to accurately estimate validity or effectiveness of HRM products and services for the criteria assessed by ratings. Using intrarater reliabilities in corrections is inappropriate and results in systematic undercorrections (i.e., systematic downward biases). If the goal is to build a cumulative HRM around rater idiosyncrasies, corrections using alpha coefficients are appropriate.

Unfortunately, appropriate reliability estimates are not routinely reported in our journals and research reports. Most published studies on job performance report internal consistency estimates such as alpha coefficients. A review of published articles since 1996, the publication year of Viswesvaran et al.'s (1996) comprehensive meta-analysis on the topic, located 12 studies reporting interrater reliabilities for the different job performance dimensions rated by supervisors. This is dismal. During this period, a total of 6864 articles appeared in *Journal of Applied Psychology*, *Personnel Psychology*, *Academy of Management Journal*, *Journal of Management*, and *International Journal of Selection and Assessment*. Of these, 907 related to job performance, contextual performance, or organizational citizenship behaviors, criteria typically assessed using ratings. We find it extremely disappointing that only about one in a thousand studies reported interrater reliability. Values are not reported in technical reports and unpublished sources either. Recently, one of the authors of this manuscript gathered over 800 criterion-related validity studies for high fidelity simulation games used for employee selection, mostly from vendors of such assessments. Not one study reported an interrater reliability coefficient for ratings criteria.

If we want to generalize our inferences across raters (e.g., supervisors rating performance), we need estimates of interrater reliability (Ones, Viswesvaran, & Schmidt, 2008; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran, Ones, Schmidt, Le, & Oh, 2014). Idiosyncratic views of individual raters, while potentially relevant for administrative decision making, cannot be the basis of a cumulative science of HRM.

Given the costs, time, effort for data collection, it is understandable that researchers and practitioners may not typically be collecting data that they do not perceive as manifestly useful for their research and practice. However, the biasing effects of

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

measurement error for ratings is large. Consequentially, effect sizes in research studies are systematically underestimated; rater idiosyncrasies constitute bases of some theories and applications; and HRM practitioners undersell the effectiveness of their products and services. Addressing the scarcity of reported interrater reliability coefficients in the HRM published and unpublished requires enhancing the knowledge, skills, and motivation of researchers and practitioners. To address knowledge and skill deficiencies, *we propose graduate training to include specific methods modules devoted to the topic*. To inform the profession at large, *we propose continuing education workshops and practitioner oriented white papers (e.g., SHRM-SIOP sponsored) on how not collecting data to assess inter-rater reliability of supervisory ratings grossly underestimates the validity (and subsequently the utility) of HRM products/services*. Practitioners can be motivated to collect and report interrater reliability by helping them understand that their products and services perform better than their measurement error attenuated, observed correlations and effect sizes indicate.

4.2. Appropriate range restriction corrections in meta-analyses

Range restriction is a complex statistical artifact. Unlike other statistical artifacts, which are caused by a specific study imperfection and are independent of other statistical artifacts, range restriction has dependencies on various other statistical artifacts (most prominently measurement error) and can take various forms depending on the type of data censuring in operation (e.g., direct range restriction, indirect range restriction).

In a series of articles, Schmidt and colleagues presented the various range restriction scenarios that operate in HRM studies, as well as methods to accurately correct for them (Hunter, Schmidt, & Le, 2006; Le & Schmidt, 2006). Applications of these methods resulted in the conclusion that cognitive measures used in employee selection and admission decisions have greater validity than previously believed (Hunter et al., 2006; Oh, Schmidt, Shaffer, & Le, 2008; Schmidt, Oh, & Le, 2006; Schmidt, Shaffer, & Oh, 2008).

With the exception of select few meta-analyses conducted to estimate criterion-related validities of measures and procedures used in selection, few meta-analyses of HRM interventions correct for range restriction. In cases where attraction, selection, and attrition produce more homogenous pools of individuals included in studies, effects of range restriction, particularly indirect range restriction, need to be considered. It is conceivable that even in studying the effectiveness of HRM interventions, researchers may need to carefully consider and correct for range restriction using appropriate methodologies. For example, a training program to enhance emotional intelligence in employees (Van Rooy & Viswesvaran, 2004) may systematically exclude participants extremely low on emotional intelligence. The important point here is that, for accurate effect size estimation in HRM, meta-analysts should correct for the appropriate types of range restriction and should carefully consider the population to which generalization of estimated relationships is desired.

5. Conclusions

Psychometric meta-analysis (Schmidt & Hunter, 2014) has enabled researchers to cumulate the voluminous empirical research in HR interventions. We now have a cumulative science and practice of HR interventions. The science and practice of HRM has benefitted from the multiple psychometric meta-analyses that have been reported over the past three decades (see DeGeest & Schmidt, 2011, for an overview). Cumulative knowledge and evidence based HR interventions require psychometric meta-analyses (Viswesvaran et al., 2014).

In this paper, we highlighted recent developments that bear on the veracity and usefulness of HRM meta-analyses. We described the necessity for meta-analyses to incorporate all available research on the focal research question as well as the barriers to inclusion of data from publicly available sources as well as practitioners. We offered some suggestions that can increase inclusiveness of meta-analytic databases. We especially focused on the issue of incorporating practitioner data into HRM meta-analyses and provided specific proposals for various stakeholders, during different stages of the meta-analytic enterprise. We certainly do not claim to have all the answers, but we hope that by encouraging thoughtfulness, providing some starting points for discussion, particularly with regard to inclusion of practitioner and consultancy studies and data, this paper helps move the field forward. HRM is an applied field. We cannot afford to ignore any research and data in building our cumulative science, particularly that generated in real organizational settings.

The invention of psychometric meta-analysis and its methods for correcting for the biasing effects of statistical artifacts was a giant leap for HRM almost 40 years ago. We must be vigilant that the meta-analyses conducted in our field use comprehensive data to combat biased conclusions and use appropriate unreliability and range restriction correction to accurately estimate effects. Psychometric meta-analysis has played a central role in our field and we hope some of the issues raised here promote thoughtful applications of this tool to advance our science and practice.

References

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, *37*(1), 5–38. http://dx.doi.org/10.1177/0149206310377113.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.*

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. http://dx.doi.org/10.1177/1745691612459060.

Bartlett, T. (2012, April 17). *Is psychology about to come undone?* Retrieved from http://chronicle.com/blogs/percolator/is-psychology-about-to-come-undone/29045

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors. *Strategic Management Journal*, *37*(2), 257–261. http://dx.doi.org/10.1002/smj.2477.

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

Bosco, F. A., Uggerslev, K. L., & Steel, P. (2016). metaBUS as a vehicle for facilitating meta-analysis. *Human Resource Management Review*. http://dx.doi.org/10.1016/j.hrmr.2016.09.013.
Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342 http://dx.doi.org/10/5k2.
Burt, C. (1917). *The distribution and relations of educational abilities*. London: Darling & Son.
Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology*, *65*(5), 543–558. http://dx.doi.org/10.1037/0021-9010.65.5.543.
Campbell, J. P. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology*, *43*(2), 231–239. http://dx.doi.org/10.1111/j.1744-6570.1990.tb01556.x.
Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates.
Campbell, J. P., & Wilmot, M. P. The functioning of theory in industrial, work, and organizational psychology. In N. R. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), Handbook of industrial, work and organizational psychology (2nd ed., Vol. 1). Sage; Thousand Oaks, CA, (in press).
Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153. http://dx.doi.org/10.1037/h0045186.
Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. http://dx.doi.org/10.1037/0003-066X.45.12.1304.
Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997.
Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology*, *65*(2), 221–249. http://dx.doi.org/10.1111/j.1744-6570.2012.01243.x.
DeGeest, D. S., & Schmidt, F. L. (2011). The impact of research synthesis methods on industrial–organizational psychology: The road from pessimism to optimism about cumulative knowledge. *Research Synthesis Methods*, *1*(3-4), 185–197. http://dx.doi.org/10.1002/jrsm.22.
Dekas, K. H., Bauer, T. N., Welle, B., Kurkoski, J., & Sullivan, S. (2013). Organizational citizenship behavior, version 2.0: A review and qualitative investigation of OCBs for knowledge workers at Google and beyond. *Academy of Management Perspectives*, *27*(3), 219–237. http://dx.doi.org/10.5465/amp.2011.0097.
Fetzer, M., & Tuzinski, K. (Eds.). (2013). *Simulations for personnel selection*. New York: Springer.
Fountain, L. (2014). The fallacies of fraud. *Internal Auditor*, *71*(4), 52–57.
Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*(1), 3–9. http://dx.doi.org/10.1002/asm.3150010103.
Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, *50*(6), 1346–1352. http://dx.doi.org/10.5465/AMJ.2007.28166119.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?*. Mahwah, NJ: Lawrence Erlbaum Associates.
Harris, W. G., Jones, J. W., Klion, R., Camara, W., Arnold, D. W., & Cunningham, M. R. (2012). Test publishers' perspective on "an updated meta-analysis": Comment on Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012). *Journal of Applied Psychology*, *97*(3), 531–536. http://dx.doi.org/10.1037/a0024767.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75*(5), 581–595. http://dx.doi.org/10.1037/0021-9010.75.5.581.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594–612 http://dx.doi.org/10/bt4t68.
International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), & InterAcademy Partnership (IAP) (2015). *Science international 2015: Open data in a big data world*. Paris, France: Authors.
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. http://dx.doi.org/10.1177/0956797611430953.
Kaplan, S., Bradley, J. C., Luchman, J. N., & Haynes, D. (2009). On the role of positive and negative affectivity in job performance: A meta-analytic investigation. *Journal of Applied Psychology*, *94*, 162–176. http://dx.doi.org/10.1037/a0013115.
Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology*, *6*(3), 252–268. http://dx.doi.org/10.1111/iops.12045.
King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, *65*(5), 507–516. http://dx.doi.org/10.1037/0021-9010.65.5.507.
Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142–164. http://dx.doi.org/10.1017/iop.2015.13.
Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods*, *11*(4), 416–438. http://dx.doi.org/10.1037/1082-989X.11.4.416.
Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*(1), 165–200 http://dx.doi.org/10/c9qbtd.
Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76*(3), 286–302. http://dx.doi.org/10.1080/03637750903074685.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Charlotte, NC: Addison-Wesley.
Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3, Pt.1), 151–159. http://dx.doi.org/10.1037/h0026141.
MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. http://dx.doi.org/10.1037/1082-989X.7.1.19.
Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. http://dx.doi.org/10.1177/1745691612460688.
Matlack, C. (2013, June). *24*. Bloomberg Business: Research fraud allegations trail a German B-school wunderkind Retrieved from http://www.bloomberg.com/bw/articles/2013-06-24/research-fraud-allegations-trail-a-german-b-school-wunderkind
McDaniel, M. A., Hurtz, G. M., & Donovan, J. J. (2006a). *An evaluation of publication bias in Big Five validity data*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006b). Publication bias: A case study of four test vendors. *Personnel Psychology*, *59*(4), 927–953. http://dx.doi.org/10.1111/j.1744-6570.2006.00059.x.
Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. http://dx.doi.org/10.2466/pr0.1990.66.1.195.
Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683–729. http://dx.doi.org/10.1111/j.1744-6570.2007.00089.x.
Morrison, D. E., & Henkel, R. E. (2007). *The significance test controversy: A reader*. New Brunswick, NJ: AldineTransaction.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. http://dx.doi.org/10.1037/1082-989X.5.2.241.

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

Novotney, A. (2014). Reproducing results. *Monitor on Psychology*, *45*(8), 32.
Nye, C. D., Su, R., Rounds, J. B., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science*, *7*(4), 384–403 http://dx.doi.org/10/33q.
Oh, I. -S., Schmidt, F. L., Shaffer, J. A., & Le, H. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning and Education*, *7*(4), 563–570. http://dx.doi.org/10.5465/AMLE.2008.35882196.
Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679–703. http://dx.doi.org/10.1037/0021-9010.78.4.679.
Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. *Industrial and Organizational Psychology*, *1*(2), 174–179. http://dx.doi.org/10.1111/j.1754-9434.2008.00033.x.
Ones, D. S., Dilchert, S., Deller, J., Albrecht, A. -G., Duehr, E. E., & Paulus, F. M. (2012a). Cross-cultural generalization: Using meta-analysis to test hypotheses about cultural variability. In A. M. Ryan, F. T. L. Leong, & F. L. Oswald (Eds.), *Conducting multinational research projects in organizational psychology: Challenges and opportunities* (pp. 91–122). Washington, DC: American Psychological Association.
Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2012b). Integrity tests predict counterproductive work behaviors and job performance well: Comment on Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012). *Journal of Applied Psychology*, *97*(3), 537–542. http://dx.doi.org/10.1037/a0024825.
Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, *15*(2), 199–228. http://dx.doi.org/10.1177/1094428111428356.
Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. http://dx.doi.org/10.1177/1745691612463401.
Pashler, H., & Wagenmakers, E. -J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. http://dx.doi.org/10.1177/1745691612465253.
Patsopoulos, N. A., Analatos, A. A., & Ioannidis, J. P. A. (2005). Relative citation impact of various study designs in the health sciences. *JAMA*, *293*(19), 2362–2366. http://dx.doi.org/10.1001/jama.293.19.2362.
Pollack, J. M., & McDaniel, M. A. (2008). An examination of the PreVisor™ Employment Inventory for publication bias. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, San Francisco, CA Retrieved from http://www.people.vcu.edu/~mamcdani/VCU%20web%20site/Publications/EI_Publication_Bias_SIOP_2008_Final_version.pdf
Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology*, *68*(3), 382–395 http://dx.doi.org/10/fq5thj.
Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*(3), 432–446. http://dx.doi.org/10.1037/0021-9010.76.3.432.
Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Replication, replication, replication. *Psychologist*, *25*(5), 346–348.
Roediger, H. L., III (2012, February). Psychology's woes and a partial cure: The value of replication. *Association for Psychological Science Observer*, *25*(2) Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html
Rogers, P. (2010, October). *Does testing for statistical significance encourage or discourage thoughtful data analysis?* Retrieved from http://genuineevaluation.com/does-testing-for-statistical-significance-encourage-or-discourage-thoughtful-data-analysis/
Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297–330. http://dx.doi.org/10.1111/j.1744-6570.2001.tb00094.x.
Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, *75*(3), 322–327. http://dx.doi.org/10.1037/0021-9010.75.3.322.
Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 103–125) (2nd ed.). New York: Russell Sage Foundation.
Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.
Sackett, P. R., & Yang, H. (2000). Corrections for range restriction: An extended typology. *Journal of Applied Psychology*, *85*(1), 112–118. http://dx.doi.org/10.1037/0021-9010.85.1.112.
Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*(10), 1173–1181. http://dx.doi.org/10.1037/0003-066X.47.10.1173.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. http://dx.doi.org/10.1037/1082-989X.1.2.115.
Schmidt, F. L. (2008). Meta-analysis: A constantly evolving research integration tool. *Organizational Research Methods*, *11*(1), 96–113. http://dx.doi.org/10.1177/1094428107303161.
Schmidt, F. L. (2015). History and development of the Schmidt-Hunter meta-analysis methods. *Research Synthesis Methods*, *6*(3), 232–239. http://dx.doi.org/10.1002/jrsm.1134.
Schmidt, F. L., Meta-regression: Advantages and disadvantages. *Career Development International* (in press).
Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*(5), 529–540 http://dx.doi.org/10/d69rvv.
Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199–223 http://dx.doi.org/10/fww5q4.
Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*(3), 183–198. http://dx.doi.org/10.1016/S0160-2896(99)00024-0.
Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.
Schmidt, F. L., & Oh, I. -S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or something else? *Archives of Scientific Psychology*, *4*(1), 32–37. http://dx.doi.org/10.1037/arc0000029.
Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*(4), 473–485. http://dx.doi.org/10.1037/0021-9010.61.4.473.
Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*(3), 432–439 (http://doi.org/10/dfm85k).
Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*(4), 901–912. http://dx.doi.org/10.1111/j.1744-6570.2000.tb02422.x.
Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*(2), 206–224. http://dx.doi.org/10.1037/1082-989X.8.2.206.
Schmidt, F. L., Oh, I. -S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology*, *59*(2), 281–305. http://dx.doi.org/10.1111/j.1744-6570.2006.00065.x.
Schmidt, F. L., Shaffer, J. A., & Oh, I. -S. (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*(4), 827–868. http://dx.doi.org/10.1111/j.1744-6570.2008.00132.x.
Schmidt, F. L., Oh, I. -S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97–128 (http://doi.org/10/b6d6c9).
Schwab, A., & Starbuck, W. H. (2009). Null-hypothesis significance tests in behavioral and management research: We can do better. In D. D. Bergh, & D. J. Ketchen (Eds.), *Research methodology in strategy and management*, *Vol. 5*. (pp. 29–54). http://dx.doi.org/10.1108/S1479-8387(2009)0000005002.
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*(5), 1055–1064. http://dx.doi.org/10.1037/a0023322.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. http://dx.doi.org/10.1177/0956797611417632.

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. http://dx.doi.org/10.1177/1745691615609918.
Stanek, K. C. (2014,). *Meta-analyses of personality and cognitive ability.* (Doctoral dissertation), Minneapolis, MN: University of Minnesota.
Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. http://dx.doi.org/10.1177/1745691614528518.
Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. http://dx.doi.org/10.1080/01621459.1959.10501497.
Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108–112. http://dx.doi.org/10.1080/00031305.1995.10476125.
Strong, E. K. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.
Tiffin, J. (1948). *Industrial psychology* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. http://dx.doi.org/10.1037/h0031322.
Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, *96*(6), 1167–1194 http://dx.doi.org/10/cqq6s5.
Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*(3), 499–530. http://dx.doi.org/10.1037/a0021196.
Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of Vocational Behavior*, *65*(1), 71–95. http://dx.doi.org/10.1016/S0001-8791(03)00076-9.
Verfaellie, M., & McGwin, J. (2011, December). The case of Diederik Stapel. Retrieved from http://www.apa.org/science/about/psa/2011/12/diederik-stapel.aspx
Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude Test Battery. *Journal of Applied Psychology*, *78*(6), 981–987. http://dx.doi.org/10.1037/0021-9010.78.6.981.
Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557–574 http://dx.doi.org/10/c8f68f.
Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I. -S. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Industrial and Organizational Psychology*, *7*(4), 507–518 http://dx.doi.org/10/5k5.
Whetzel, D. L. (2006). *Publication bias the validity of customer service measures.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
Wiernik, B. M. (2016). *The nomological network of classic and contemporary career preferences: meta-analyses of vocational interests and new career orientations (Doctoral dissertation).* Minneapolis, MN: University of Minnesota.
Yoon, K., Schmidt, F. L., & Ilies, R. (2002). Cross-cultural construct validity of the five-factor model of personality among Korean employees. *Journal of Cross-Cultural Psychology*, *33*(3), 217–235. http://dx.doi.org/10.1177/0022022102033003001.
Ziliak, S. T., & McCloskey, D. N. (2007). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor, MI: University of Michigan Press Retrieved from http://www.deirdremccloskey.com/articles/stats/preface_ziliak.php

Please cite this article as: Ones, D.S., et al., Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.09.011