6-24-2015

Improving the rigor of quantitative HRD research: Four recommendations in support of the general hierarchy of evidence

Kim Nimon
University of Texas at Tyler, knimon@uttyler.edu

Marina Astakhova

Improving the Rigor of Quantitative HRD Research: Four Recommendations in Support of the General Hierarchy of Evidence

Kim F. Nimon, Marina Astakhova

Keywords: meta-analysis, experimental designs, retrospective pretest, mediated designs, cross-sectional research, common method bias, common method variance, human resource development, evidence-based management, general hierarchy of evidence

Recently, in an editorial for *Human Resource Development Quarterly* (*HRDQ*) 26(2), Gubbins and Rousseau (2015) offered six avenues for human resource development (HRD) researchers to engage in translational research or research that supports evidence-based management. Drawing from Woolf (2008), they indicated that:

> Translational research is classified into two distinct domains: *T1 research* refers to the "research-to-practitioner" enterprise of translating knowledge from the basic sciences into the development of new interventions, models, guidelines, or products; and *T2 research* refers to the translation of research into practice such that new intervention/product from T1 are *used* in everyday practice and decision making.… For example, T1 HRD research might explore how to better motivate learning transfer and develop principles for practice. T2 HRD research might evaluate the effectiveness of these principles when used in real-world settings and how effectively they are applied under varying conditions. (p. 110)

To help HRD "become a field where evidence-based practice can readily take place", Gubbins and Rousseau (2015, p. 6) presented a general hierarchy of evidence based on research design to help assess the quality of research that seeks to answer questions regarding "what works" or "does X cause Y". At the

Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/hrdq.21219

top of the hierarchy were meta-analytic reviews, followed by experimental and then quasi-experimental designs, followed by longitudinal and then cross sectional or survey designs, with case studies and then expert opinions forming the base of the hierarchy. They rightly noted that experimental and quasi-experimental designs provided the strongest evidence of causality and that bias was strongest in expert opinions and weakest in meta-analytic designs.

In this editorial, we build on the work of Gubbins and Rousseau (2015) and offer four recommendations to help HRD researchers support the general hierarchy of evidence. Our recommendations correspond to the top five levels of Gubbins and Rousseau's hierarchy. First, at the top of the hierarchy, we recommend researchers report sufficient information to support meta-analytic reviews. Second, we present the retrospective pretest as an accessible approach to experimental designs. Third, we review the continuum of mediated designs. Lastly, we discuss how to make a case for and employ proper procedures when using cross-sectional survey data.

In preparation for this editorial, we reviewed articles published in *Human Resource Development Quarterly* (*HRDQ*) over the last five years (2010 – 2014) to benchmark the designs of the studies reported as well as specifics related to our recommendations. Our review found that of the 63 articles that reported quantitative research, one was a meta-analytic review, one was a citation analysis, two employed mixed-methods, five were measurement studies, nine were based on experimental/quasi-experimental designs, and the remaining 45 stemmed from cross-sectional research.

Supporting Meta-Analytic Reviews

In 2013, Ellinger, Anderson, Gubbins, Lunn, Nimon, Sheehan, and Werner advised researchers conducting quantitative studies to comply with standards (e.g., American Educational Research Association, 2006) that call for the reporting of sample demographics, descriptive statistics including correlation or covariance matrices, test statistics, and effect sizes, among others. Complying with these standards encourages meta-analytic thinking which Thompson (2002b) defined as "both (a) the prospective formulation of study expectations and design by explicitly invoking prior effect sizes and (b) the retrospective interpretation of new results, once they are in hand, via explicit, direct comparison with the prior effect sizes in the related literature" (p. 28). Observing reporting standards also supports meta-analytic reviews, as the statistics and the study features within such articles may be subsequently coded and analyzed.

Although not explicitly mentioned in Ellinger et al. (2013), reliability coefficients can be important to meta-analysis studies, as effect sizes may be attenuated when reliability is less than perfect (but see Nimon, Zientek, and Henson [2012] that illustrates how effect sizes may be inflated in the presence of correlated error), and should therefore be reported in quantitative studies that analyze scale scores composed of multiple items. When reliability coefficients are reported, meta-analytic researchers have the opportunity to adjust effect sizes based on the measurement error implied in the reliability coefficient (Schmidt, Le, & Oh, 2009). When reliability coefficients are not reported, synthesists may have to "borrow relevant coefficients from test manuals or reports of similar research" or employ procedures for integrating effect size estimates that have or have not been corrected (Shadish & Haddock, 2009, p. 260).

Note, however, that reliability is a property of scores, not the instrument, and is therefore sample and context dependent (Thompson, 2002a). Based on a study comparing sample compositions and variabilities between published studies and test manuals, Vacha-Haase, Kogan, and Thompson (2000) determined that using reliability coefficients from prior studies was "modestly plausible *only if* [italics added] the compositions and variabilities of the two samples are explicitly and directly compared" (p. 521). Instead, Vacha-Haase et al. recommended that researchers follow Wilkinson and The APA Task Force on Statistical Inference (1999) and report the reliability of their own scores, even in substantive (i.e., non-measurement) studies. As it relates to meta-analyses, indefensible score reliability inductions may present problems as results will be biased without proper correction for measurement error (cf. Schmidt, Le, & Oh, 2009).

To benchmark the degree to which articles published in *HRDQ* might be included in a meta-analysis, we reviewed applicable articles to see how many reported effect sizes (or statistics that could be used to compute effect sizes) and reliability coefficients as warranted. All of the articles that reported quantitative research (n = 63) reported an effect size, test statistic, or set of descriptive statistics that could potentially be used in a meta-analysis. However, six articles did not report a comprehensive set of reliability coefficients, despite analyzing scale scores that were comprised of responses from multiple items. For example, in the study conducted by Nimon, Zigarmi, Houson, Witt, and Diehl (2011) that provided initial evidence of construct validity for Work Cognition Inventory (WCI) scores, reliability coefficients were only provided for WCI scale scores, even though the authors reported correlations to scores from the Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1998); Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), and Marlow-Crowne Social Desirability Scale (MSCD; Crowne & Marlowe, 1960). In another instance, Luthans, Avey, Avolio, and Peterson (2010) indicated that subscales "demonstrated reliability alphas greater than 0.70 with the exception of resilience" (p. 53). Similary, Neiminen, Smerek, Kotrba, and Denison (2013) reported a range of reliability coefficients (i.e., > 0.90) for the scores analyzed in their study. While such ranges demonstrate the level of reliability of data at certain benchmarks, they do not indicate the precise level of measurement error which is needed when comparing effect sizes across multiple studies. Note that only 5% of the applicable articles reviewed did not report reliability. This is in stark contrast to

Vacha-Haase and Thompson (2011) who found that across 47 reliability generalization meta-analysis studies which represented 12,994 primary reports, 54.6% did not mention reliability. Although more work within *HRDQ* will help assure that all applicable articles report sufficient information to be considered in meta-analytic reviews, the current findings suggest that *HRDQ is* laying a foundation for future meta-analytic reviews, even if few such studies have been published recently.

Retrospective Approach to Experimental/ Quasi-Experimental/Pre-Experimental Designs

Campbell and Stanley's (1963) seminal book on experimental and quasi-experimental designs for research offered the retrospective pretest as an extension to the pre-experimental design where two groups (one which has experienced a treatment, intervention, or condition [i.e., X] and one which has not) are measured. Although Campbell and Stanley did not provide a definition for the retrospective pretest, it is typically understood that the retrospective pretest (also called the thentest) is given after an intervention and asks participants to assess their pre-intervention knowledge, skills, or attitudes. In describing the retrospective pretest approach, Campbell and Stanley reviewed two studies (i.e., Deutsch & Collins, 1951; Information and Education Division, 1947) to demonstrate the advantage of including a retrospective pretest as a practical means to assess if there are pretest differences between groups and to rule out plausible rival hypotheses associated with the static-group comparison design (i.e., Design 3).

In the Information and Education Division (1947) study, comparisons between the attitudes of whites assigned to a racially mixed combat infantry unit versus an all-white unit were of causal interest. In a "posttest" interview, participants were asked to indicate their present attitudes towards Negroes as well as to retrospectively recall their attitudes prior to the assignment. Results indicated that while there were no differences between the two groups based on the retrospective pretest accounts, the whites assigned to the racially mixed unit had more favorable attitudes toward Negroes, "thus increasing the plausibility that prior to the assignment there had been no difference" (Campbell & Stanley, 1963 p. 66).

Campbell and Stanley (1963) reviewed the results of a similar study by Deutsch and Collins (1951) which found that housing project occupants in integrated units had more favorable attitudes than their counterparts in segregated units. As noted by Campbell and Stanley:

> Having only posttest measures, the differences they found might have been regarded as reflecting selection biases in initial attitudes. The interpretation that the integrated experience *caused* [italics added] the more variable attitudes was enhanced when a retrospective pretest showed

> no differences between the two types of housing groups in remembered prior attitudes. (p. 66)

A little over one decade after Campbell and Stanley's (1963) seminal publication, Howard, Ralph, Gulanick, Maxwell, Nance, and Gerber (1979) proposed extending pretest-posttest designs by *adding* the retrospective pretest to moderate the confounding effect of experience limitation. Through a series of studies, Howard et al. concluded that when a response shift occurred, retrospective pretest accounts were more valid than traditional pretest accounts, where a response shift was identified as a statistically and practically significant difference between retrospective and traditional pretest scores. In the first study, Howard et al. assessed male non-commissioned officers before and after a communication skills training workshop designed to reduce dogmatism. Rather than finding a *decrease* in self-report levels of dogmatism, they found that participants reported being *more* dogmatic after the workshop. Posthoc interviews revealed that participants changed their initial perceptions of their level of dogmatism as a result of the workshop which helped explain the paradoxical findings. In the second study, Howard et al. measured change by means of a traditional pretest-posttest design and a retrospective pretest-posttest design and found radically different results. The remainder of the three studies found support for the retrospective pretest-posttest design by correlating indices of change based on self-report to objective measures of change. Howard et al. concluded that the use of pretest, posttest, and retrospective pretest provided "a more sensitive, assessment of a subject's perspective of personal change" and "another valuable dimension to evaluation research endeavors" (p. 22).

In a little over two decades past the seminal work of Howard et al. (1979), evaluators (e.g., Lamb & Tschillard, 2005; Martineua, 2004; Raidl, Johnson, Gardiner, Denham, Spain, & Lanting, 2004), suggested *replacing* the traditional pretest in pretest-posttest designs with the retrospective pretest as a practical and valid means to determine program outcomes. While replacing the traditional pretest with the retrospective pretest does not allow for the assessment of response shift, the design was expected to mitigate the effects of pretest sensitization, maturity, and mortality, based on prior research. However, the vast amount of contemporary studies using the retrospective pretest in lieu of the traditional pretest were based on survey designs where post and retrospective pretest items were placed on the same survey either side by side or one underneath another, calling into question biases associated with implicit theories of change or stability. In response, Nimon, Zigarmi, and Allen (2011) tested the validity of retrospective pretest measures across four designs and found that administering posttests separately from retrospective pretests produced more valid results than placing posttest and retrospective pretest items on the same survey.

In our aforementioned review of 63 quantitative articles, we found that 14% were based on some form of experimental design. Given that most of the definitions of HRD (e.g., McLagan, 1989) consider training and development, it is somewhat surprising that so few articles report on pre-, quasi-, or experimental designs that assess knowledge, skills, or attitudes (KSAs) either after or before and after a training or development intervention. We recognize that traditional pretest-posttest designs are pragmatically difficult as study participants may arrive late to or leave early from an intervention thereby creating small effective sample sizes. In addition, participants may not know what they do not know at the onset of an intervention, thereby creating a response shift in their pretest and posttest responses which may result in attenuated effect sizes (Nimon, 2014).

We, therefore, recommend that HRD researchers consider the retrospective pretest as a means to refine their evaluation designs. The retrospective pretest may mitigate the bias associated with response-shift, as participants have the opportunity to evaluate their pre- and post-intervention knowledge, skills, or attitudes using the same frame of reference (Howard et al., 1979). Depending on the encompassing design, the retrospective pretest may also mitigate the effects of maturity, mortality, and pretest-sensitization. Limitations of the retrospective pretest include biases associated with implicit theories of change or stability, impression management, and memory distortion (Hill & Betz, 2005; Nimon, 2014). However, such biases may be mitigated by proper survey administration and procedures. Recommendations for a rigorous design incorporating the retrospective pretest also include incorporating a control measure, administering the posttest separately from the retrospective pretest, and allowing sufficient time between the administration of the posttest and retrospective pretest such that participants cannot artificially communicate change or stability in order to please the facilitator of the intervention (Nimon et al., 2011).

Continuum of Mediated Designs

Theory-building research "can help the HRD profession address the call for HRD theory, offer a means for stepping up to the potential problems in HRD practice [...] and provide methods for reducing the incidence of practice based on incomplete espoused theories" (Lynham, 2000, p. 159). Theory building often explains *why* variables are related. Although providing conceptual justification for *why*-mechanisms is important, "theory alone isn't enough" and why-mechanisms (i.e., mediating processes) need to be supported empirically (Bono & McNamara, 2011, p. 659). To execute a solid and compelling mediation study, careful attention to the study design is warranted, as an ounce of prevention may be worth a pound of cure later. Important issues that may pose a challenge for a mediation study include a) choice of mediators, b) number of formal hypotheses statements, c) strategies in collecting data on the independent variable (X), mediator (M) and dependent variable (Y), and d) statistical analyses to test mediation hypotheses. We discuss these issues in more detail below.

Our aforementioned review of quantitative studies revealed a total of 23 studies that hypothesized and tested mediation models. Of 23 studies, 13 studies used a single-mediator model (e.g., Halbesleben & Stoutner, 2013; Moon, Choi, & Jung, 2012), 9 studies tested a multiple-mediator model (e.g., Froehlich, Segers, & Van den Bossche, 2014; Sommer & Kulkarni, 2012) and one study employed a moderated mediation design (Madera et al., 2011). Although a single-source cross-sectional design persists in testing mediation, several studies used multi-source data samples (e.g., [employee-supervisor dyads], Kang & Bartlett, 2013; [customer-service provider dyads], Halbesleben & Stoutner, 2013) or employed an experimental (Sookhai & Budworth, 2010) or longitudinal designs (Madera et al., 2011). Finally, we found that HRD researchers used diverse analytic methods to test mediation (e.g., hierarchical regression, structural equation modeling [SEM], Sobel test, bootstrapping, or decomposition analysis), which, in most part, reflect the state-of-the-art methods commonly used in industrial psychology and management (Baron & Kenny, 1986; Preacher & Hayes, 2004; 2008). Overall, our review encourages the active engagement of HRD scholars in a theory-building discourse that attempts to theorize and empirically test *why* relationships between variables hold. To maintain the positive momentum in HRD research, we offer suggestions on how to further increase the rigor and avoid common pitfalls in mediation design.

Choice of Mediators

It seems straightforward and perhaps a little mundane to caution the reader that choice of mediators should be driven by theory. It is more challenging to decide when mediators should be included in the model and whether a single- or multiple-mediator model is preferred. Bono and McNamara (2011) suggested that when an area of inquiry is new, the focus should first be on establishing a causal relationship between X and Y. Once the causality is established, it becomes essential to explain how or why the causal effect occurs (i.e., propose and test a mediating process). For example, past international experience and cross-cultural training have long been established in HRD research as antecedents of expatriates' adjustment (for a review, see Bhaskar-Shrinivas, Harrison, Shaffer, & Luk, 2005). Moon et al. (2012) extended this knowledge by proposing that mediating effect of cultural on the relationships between international experience and cross-cultural training and adjustment levels.

As an area of inquiry matures, multiple- versus single-mediation models may better explain the phenomena, because such models provide a more accurate assessment of mediation effects (Bono & McNamara, 2011; MacKinnon, 2000). The caveat is not to create an array of unrelated mediators but rather include only those that are conceptually linked. For example, researchers may have a difficult time convincing reviewers that employees' organizational tenure, self-efficacy, and engagement are multiple mediators of the relationship between perceived support for participation in HRD practices and intention to turnover. In contrast, the mediating effects of three foci of engagement (cognitive, emotional, and behavioral) on the same relationships well exemplify the use of conceptually related mediators (Shuck, Twyford, Reio, & Shuck, 2014). Another way of "marrying" multiple mediators is by proposing a multiple sequential mediator model. For example, Sommer and Kulkarni (2012) found that the link between constructive feedback and organizational citizenship behavior intentions is complex and is transmitted through a two-stage mediation process that includes perceived respect and then positive and negative affect. Similarly, the researchers found a sequential mediation path between constructive feedback and job satisfaction which goes through perceived respect or opportunities for advancement and then positive or negative affect.

Number of Formal Hypotheses Statements

The number of formal hypotheses statements in a mediation study typically varies and may range from one to four. A common approach is to advance four formal hypotheses which would mirror the four steps popularized by Baron and Kenny (1986): the relationships between X and Y, X and M, M and Y, and an indirect (mediating) effect (an effect of the X on Y, while controlling for M). Because an indirect effect is still plausible in the absence of an association between X and Y (Hayes, 2009; Mathieu & Taylor, 2006), some studies offer three (instead of four) hypotheses (no formal hypothesis for the relationship between X and Y) (e.g., Sommer & Kulkarni, 2012). Other studies advanced only two hypotheses: an association between X and Y and a mediating effect (e.g., Yamkovenko & Holton, 2010). Finally, researchers may limit the number of formal hypotheses statements to only one. Does a magic number of hypotheses exist? The answer is "no" as the number of hypotheses remains a researcher's choice. However, researchers need to provide explicit theoretical arguments to each of the mediation steps (even in the absence of formal hypotheses for those steps), because a mediating effect can only be justified if the relationships between X and M and M and Y are supported. Furthermore, researchers should remember that, although the relationships between X and M and M and Y are prerequisites for a mediating effect, a case of a mediating process cannot be assumed merely "by extension" from these two relationships. In other words, if a researcher proposes X→M and M→Y, it is insufficient to immediately conclude that M is a mediator. An additional argument should be made to explain why M is expected to transmit the link between X to Y. For example, Kang and Bartlett (2013), who proposed a mediating effect of psychological empowerment on the relationship between perceived external prestige and customer-oriented citizenship behaviors, argued that the conceptualization of empowerment as an intrinsic value suggests that individuals will likely internalize their perceptions of organizational prestige before performing citizenship behaviors. The absence of such theorizing for the last mediation step will signal an incomplete hypotheses development to reviewers.

Strategies in Collecting Data on X, M, and Y

The nature of a mediating process assumes a time-lag between X and Y, suggesting the need for experimental or longitudinal designs. In practice, though, the application of such designs among HRD researchers remains infrequent. Our review revealed that, out of 23 mediation studies, only two studies used primary (Madera et al., 2011) or secondary longitudinal data (Park & Jacobs, 2011), two studies employed experiments (Hui, Sue-Chan, & Wood, 2013; Sookhai & Budworth, 2010) and the rest employed cross-sectional samples.

What are the consequences of a cross-sectional, or *one-shot or opportunistic* study design, for a mediation study? Results of such a study will most likely be biased, as researchers may find indirect effects when only direct effects exist and find direct effects when only indirect effects exist (Maxwell, Cole, & Mitchell, 2011). Is there any cure for such fallacies? The best advice is to avoid one-shot study designs but rather move up the continuum of mediated designs. If a longitudinal or an experimental study is not possible, a *sequential* study design can be implemented. Researchers can conduct a three-stage study and collect data on X at Time 1, on M at Time 2, and on Y at Time 3. Although each study in a series may have flaws, "together the studies may allow for stronger inferences and more generalizable results than would any single study on its own" (Bono & McNamara, 2011, p. 660). However, despite the benefits of the time lag, this design has limitations. As Hoyle and Robinson (2004) noted, the effects of a predictor at Time 1 on an outcome at Time 2 may not be isolated from the same outcome at Time 1.

For example, in the mediation model in which harmonious job passion (Time 1) translates into employee engagement (Time 3) via job satisfaction (Time 2), it is not clear whether the path between job passion and engagement is a true representation of the association between the two constructs, or whether it reflects some stable timeless association between job passion and engagement (employees with job passion are always engaged). Similarly, it is unclear whether job passion (Time 1) and job satisfaction (Time 2) are indeed isolated (wouldn't passionate employees be always job satisfied?). One solution to the "inferential conundrum" of one-shot and sequential strategy designs (Hoyle & Robinson, p. 223) is a *replicative* approach, in which the predictor and the outcome are measured at both points of time. For example, a researcher would measure job passion and employee engagement at Time 1 and would replicate those measurements at Time 2. Controlling for such hard-to-isolate effects among the variables would facilitate the assessment of a true (unique) variability in employee engagement (Time 2) due to job passion (Time 1). Moving up the ladder of the mediation designs will likely reduce validity threats and allow for “persuasive tests of causal hypotheses by ruling out alternatives that undermine causal inferences” (Hoyle & Robinson, p. 223).

Statistical Analyses to Test Mediation

The seminal work of Baron and Kenny (1986) advanced a four-step regression test of mediation accompanied by the Sobel test to assess significance of a mediating effect. Although Baron and Kenny’s approach has been dominant for decades, over the years, methods used to test mediating models “have grown in sophistication” (Hayes, 2009, p. 408). An example includes the use of structural equation modeling (SEM) which allows researchers to control for measurement error and permits alternative model testing to eliminate alternative explanations for the hypothesized relationships. Alternatively, researchers can use SPSS and SAS macros developed by Preacher and Hayes (2004; 2008) which combine the best features of the traditional Baron and Kenny’s approach with bootstrapping, thus allowing for a direct test of the significance of mediating effects with a simple command. Bootstrapping used to test the significance of an indirect effect is superior to Sobel test because it does not depend on multivariate normal data or a known sampling distribution (Preacher & Hayes, 2004).

Despite the rise of advanced statistical methods for testing mediating processes, the approaches used in some of the *HRDQ* articles in the last five years do not seem to have kept pace with statistical advances. A few researchers continue using the traditional Baron and Kenny’s approach and Sobel test (e.g., Hui et al., 2013; Moon et al., 2012; Sookhai & Budworth, 2010). Although the application of SEM to assess mediating processes is on the rise among HRD scholars, only a few studies have used SEM to address possible alternative explanations within the mediation model (Halbesleben & Stoutner, 2013; Kang & Bartlett, 2013). Even fewer studies have tested the significance of indirect effects via bootstrapping effect decomposition (see Gillet & Vandenberghe, 2014 and Song, Kolb, Lee, & Kim, 2012, for exceptions). Finally, only one out of 23 studies (Walsh, Bauerle, & Magley, 2013) used the Preacher and Hayes’ (2004) SPSS macros and only one study (Morris, Messal, & Meriac, 2013) applied a novel phantom modeling approach to test the significance of mediating effects (Macho & Ledermann, 2011). Phantom approach typically suits multiple mediator situations and involves a series of paths constrained to specific values to calculate the effect of each mediator separately. Given a variety of advanced methods to test mediation, we encourage HRD scholars to take advantage of them.

The Case and Employing Proper Procedures for Cross-sectional Survey Designs

The major savings in time and cost of cross-sectional data sampling make it an attractive alternative to longitudinal and experimental studies (Maxwell et al., 2011). Yet, these savings often come at the expense of a desk-rejection. According to Bono and McNamara (2011), "rejection does not happen because such data are inherently flawed or because reviewers or editors are biased against such data" but because many research questions address issues of causality or change which can only be addressed by longitudinal, experimental and panel data and not via a cross-sectional study (p. 657). A majority of HRD studies involve issues of causality or change. For example, Joo, Jeung, and Yoon (2010) examined the influences of core self-evaluations, job autonomy, and intrinsic motivation on in-role job performance. Trudel and Reio Jr. (2011) tested whether conflict management styles will have an effect on workplace incivility.

What if a cross-sectional study is the only choice for a researcher? Although we strongly encourage HRD scholars to pursue higher level designs such as meta-analysis, randomized control studies, longitudinal or experimental studies (Gubbins & Rousseau, 2015), we nevertheless provide suggestions on how to make an appealing case when using a cross-sectional design. Attention to a matching research question and common method bias (CMB) may reduce reviewers' concerns.

Matching Research Questions

It is well known that "matching research design to research questions is as much art as science" (Bono & McNamara, 2011, p. 657). This implies that no causal relationships can be inferred from a cross-sectional study. As such, HRD scholars should not succumb to the temptation of using the words "increases/decreases," "influences," "affects," changes" or "causes" in cross-sectional research. Instead, the preferred vocabulary should include "correlates," "is related to" or "is associated with." For example, in a cross-sectional study that assessed job efficacy and job satisfaction, one can only propose and test a positive/negative association between the two constructs.

Common Method Variance and Common Method Bias

Common method variance (CMV) is commonly defined as variance attributable to the measurement method rather than to the construct of interest (Fiske, 1982; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Common method variance is one of the major sources of systematic measurement error (Podsakoff et al., 2003). It is important to differentiate between CMV and common method bias (CMB). While CMV indicates that variance in observed scores is partly attributable to the effect of a measurement method, CMB points out the degree to which a methods effect inflates correlations.

Although CMB is shown to present a validity threat for different measures and in different study contexts (Cote & Buckley, 1987), cross-sectional designs are particularly vulnerable to the inflation of correlations due to CMB (Lindell & Whitney, 2001). However, the situation is not hopeless. Indeed, it is almost unrealistic to conduct a perfect flawless study that has absolutely no threat of CMB. The goal should be to reduce the likelihood of CMB, when possible. We do not attempt to repeat the content from seminal literature about controlling common method bias (e.g., Conway & Lance, 2010; Cote & Buckley, 1987; Lindell & Whitney, 2001; Podsakoff et al. 2003; Spector, 1987; Williams & Brown, 1994). Instead, we intend to increase awareness of CMB remedies among HRD researchers and encourage their regular application.

Other-reports

Although self-report measures have received substantial criticism in the literature, they remain appropriate for certain constructs (Conway & Lance, 2010). For example, scores on job satisfaction, core self-evaluations or perceptions of organizational support are gathered via self-reports. However, for constructs such as job performance or organizational citizenship, supervisor ratings are superior. An illustration of the proper use of self- versus other-reports is the study by Kang and Bartlett (2013) that examined the mediating role of psychological empowerment on the relationship between perceived role prestige and customer-oriented citizenship behaviors. While the measures for the first two constructs were self-reported, customer-oriented citizenship behaviors were assessed by supervisors. Other-reports may include but are not limited to responses of subordinates, co-workers, customers, etc.

Controlling for Common Method Bias

Podsakoff et al. (2003) popularized a wide range of procedural approaches that can proactively address CMB. For example, protecting respondents' anonymity during data collection may reduce evaluation apprehension. An intentional ordering of survey questions so as to capture the dependent variables first and having a survey question or scales that are unrelated to the particular study may provide psychological separation of the independent and dependent variables. Other approaches include the use of other-measures, filter questions, or valid measurement scales. Our five-year review demonstrated that procedural approaches to mitigate CMB are relatively infrequent among HRD scholars and tend to focus on protecting respondents' anonymity and using time separation of responses. For example, Ghosh, Reio, and Haynes (2012) "took a more procedural approach, assuring participant anonymity and that there was no right or wrong answers" (p. 50). Morris et al. (2013) also assured anonymity of respondents and collected predictor measures at a time separate from the criteria measure. Because procedural remedies are relatively easy steps to implement, we encourage HRD scholars to use and report them more habitually.

Testing for Common Method Bias

Our review demonstrates that statistical remedies for CMB are even less common than procedural remedies. The majority of researchers who do test for

CMB continue applying Harman's single-factor test, which has long been criticized for doing "nothing to statistically control (or partial out) method effects," as it is highly unlikely that a single factor will emerge in data (Podsakoff et al., 2003). Podsakoff et al. (2003) described various statistical remedies for CMB, such as controlling for the effects of a directly measured or unmeasured latent methods factor, multiple method factors, or partial correlation techniques (i.e., the use of a marker variable). Although these techniques are not free of criticism, they provide a more accurate assessment of CMB and would therefore result in less biased outcomes. The application of such approaches will help HRD researchers quell common source bias concerns more effectively, resulting in more valid study outcomes.

Concluding Comments

Building on Gubbins and Rousseau's (2015) general hierarchy of evidence, we offer four recommendations to HRD researchers: (a) report sufficient information to support meta-analytic reviews, (b) consider integrating the retrospective pretest into experimental designs, (c) move up the continuum of mediated designs, (d) make a case for and employ proper procedures when using cross-sectional survey data. Like our colleagues, Gubbins and Rousseau, we would be remiss if we did not also recommend that researchers conduct research on "important 'what works' questions. Although it is necessary to recognize that research methods (cf. Gubbins & Rousseau, 2015; Murnane & Willett, 2010), are important, even the most optimally designed research study may not be impactful or make a difference in the fields of HRD or management.

Consider the research on employee engagement, for example. Saks and Gruman (2014) remarked that although research on employee engagement has been flourishing over the past decade, there has not been "enough attention to the things that really matter: meaning, measurement, and theory. The frenzy of research has left many important questions unanswered. As a result we do not know what causes employee engagement, the effect of employee engagement on employee and organizational outcomes, and the most effective program and interventions for improving employee engagement" (p. 178).

We encourage researchers to engage in some risk taking and spend less time on the low-hanging fruit of convenient research and more time on research that matters. Many scholars and scholar-practitioners may have "bucket lists" for their personal lives (e.g., hiking to Machu Picchu, visiting the seven wonders of the ancient world, going to all of the state fairs). Might scholars and scholar-practitioners have "bucket lists" for their professional lives? A research bucket list might include studies that seek to answer some of the questions posed by Saks and Gruman. So consider this: What is on your research bucket list and how will you get started? We hope that this editorial provides you with some fodder for further developing and more rigorously implementing your research bucket list and that conversations for improving the rigor of quantitative HRD research will continue both in the published literature as well as at future conferences.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality & Social Psychology*, *51*(6), 1173–1182.

Bhaskar-Shrinivas, P., Harrison, D. A., Shaffer, M. A., & Luk, D. M. (2005). Input-based and time-based models of international adjustment: Meta-analytic evidence and theoretical extensions. *Academy of Management Journal, 48*(2), 259–281.

Bono, J. E., & McNamara, G. (2011). From the editors: Publishing in AMJ-Part 2: Research design. *Academy of Management Journal, 54*, 657–660.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: RandMcNally.

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology, 25*, 325–334.

Cote, J. A., & Buckley, R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research*, *24*, 315–318.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354.

Deutsch, M., & Collins, M. E. (1951). *Interracial housing: A psychological evaluation of a social experiment*. Minneapolis. MN: University of Minnesota Press.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49,* 71–75.

Ellinger, A. D., Anderson, V., Gubbins, C., Lunn, M. L., Nimon, K. F., Sheehan, M., & Werner, J. M. (2013). The generous spirit of the peer review process: Perspectives and insights from the HRDQ Editorial team on providing high-quality reviews. *Human Resource Development Quarterly*, *24*, 417–428.

Fiske, D. W. (1982). Convergent–discriminant validation in measurements and research strategies. In D. Brinbirg & L. H. Kidder (Eds.), *Forms of validity in research* (pp. 77–92). San Francisco, CA: Jossey-Bass.

Froehlich, D., Segers, M., & Van den Bossche, P. (2014). Informal workplace learning in Austrian banks: The influence of learning approach, leadership style, and organizational learning culture on managers' learning outcomes. *Human Resource Development Quarterly*, *25*, 29–57.

Ghosh, R., Reio, T. G., & Haynes, R. K. (2012). Mentoring and organizational citizenship behavior: Estimating the mediating effects of organization-based self-esteem and affective commitment. *Human Resource Development Quarterly*, *23*, 41–63.

Gillet, N., & Vandenberghe, C. (2014). Transformational leadership and organizational commitment: The mediating role of job characteristics. *Human Resource Development Quarterly*, *25*, 321–347.

Gubbins, C., & Rousseau, D. M. (2015). Embracing translational HRD research for evidence-based management: Let's talk about how to bridge the research–practice gap. *Human Resource Development Quarterly, 25*, 109–125.

Halbesleben, J. R. B., & Stoutner, O. K. (2013). Developing customers as partial employees: Predictors and outcomes of customer performance in a services context. *Human Resource Development Quarterly*, *24*, 313–335.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, *76*, 408–420.
Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26*, 501–507.
Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal validity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*, 1–23.
Hoyle, R. H., & Robinson, J. C. (2004). Mediated and moderated effects in social psychological research. In C. Sansone, C. C. Morg, & A.T. Panter (Eds.), *Handbook of methods in social psychology* (pp. 213–233). Thousand Oaks, CA: Sage.
Hui, R. T. Y., Sue-Chan, C., & Wood, R. E. (2013). The contrasting effects of coaching style on task performance: The mediating roles of subjective task complexity and self-set goal. *Human Resource Development Quarterly*, *24,* 429–458.
Information and Education Division, U. S. War Department (1947). Opinions about Negro infantry platoons in white companies of seven divisions. In T. M. Newcomb & E. L. Hartley (Eds.), *Readings in social psychology* (pp. 542–546). New York: Holt.
Joo, B., Jeung, C., & Yoon, H. J. (2010). Investigating the influences of core self-evaluations, job autonomy, and intrinsic motivation on in-role job performance. *Human Resource Development Quarterly*, *21*, 353–371.
Kang, D-s., & Bartlett, K. R. (2013). The role of perceived external prestige in predicting customer-oriented citizenship behaviors. *Human Resource Development Quarterly*, *24*, 285–312.
Lamb, T. A., & Tschillard, R. (2005, Spring). Evaluating learning in professional development workshops: Using the retrospective pretest. *Journal of Research in Professional Learning.* Retrieved October 19, 2005, from http://www.nsdc.org/library/publications/research/lamb.pdf
Lindell, M. K., & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology*, *86*, 114–121.
Luthans, F., Avey, J. B., Avolio, B. J., & Peterson, S. J. (2010). The development and resulting performance impact of positive psychological capital. *Human resource development quarterly*, *21*, 41–67.
Lynham, S. A. (2000). Theory building in the human resource development profession. *Human Resource Development Quarterly, 11,* 159–178.
Macho, S., & Ledermann, T. (2011). Estimating, testing, and comparing specific effects in structural equation models: The phantom model approach. *Psychological Methods*, *16*, 34–43.
MacKinnon, D. P. (2000). Contrasts in multiple mediator models. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research: New methods for new questions* (pp. 141–160). Mahwah, NJ: Erlbaum.
Madera, J. M., Steele, S. T., & Beier, M. (2011). The temporal effect of training utility perceptions on adopting a trained method: The role of perceived organizational support. *Human Resource Development Quarterly*, *22*, 69–86.
Martineua, J. (2004). *Evaluating leadership development programs: A professional guide.* Greensboro, NC: Center for Creative Leadership.
Mathieu, J. E., & Taylor, S. R. (2006). Clarifying conditions and decision points for mediational type inferences in Organizational Behavior. *Journal of Organizational Behavior*, *27*, 1031–1056.
Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross- sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, *46*, 816–841.
McLagan, P. A. (1989). *Models for HRD practice.* Alexandria, VA: American Society for Training and Development.
Moon, H. K., Choi, B. K., & Jung, J. S. (2012). Previous international experience, cross-cultural training, and expatriates' cross-cultural adjustment: Effects of cultural intelligence and goal orientation. *Human Resource Development Quarterly*, *23*, 285–330.
Morris, M. L., & Messal, C. B., & Meriac, J. P. (2013). Core self-evaluation and goal orientation: Understanding work stress. *Human Resource Development Quarterly*, *24*, 35–62.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford, England: Oxford University Press.

Nieminen, L. R., Smerek, R., Kotrba, L., & Denison, D. (2013). What does an executive coaching intervention add beyond facilitated multisource feedback? Effects on leader self-ratings and perceived effectiveness. *Human Resource Development Quarterly*, *24*, 145–176.

Nimon, K. (2014). Explaining differences between retrospective and traditional pretest self-assessments: Competing theories and empirical evidence. *International Journal of Research & Method in Education, 37,* 256–269.

Nimon, K., Zientek, L. R., & Henson, R. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3*(102), 1–13.

Nimon, K., Zigarmi, D., & Allen, J. (2011). Measures of program effectiveness based on retrospective pretest data: Are all created equal? *American Journal of Evaluation, 32,* 8–28.

Nimon, K., Zigarmi, D., Houson, D., Witt, D., & Diehl, J. (2011). The work cognition inventory: Initial evidence of construct validity. *Human Resource Development Quarterly*, *22*, 7–35.

Park, Y. J., & Jacobs, R. L. (2011). The influence of investment in workplace learning on learning outcomes and organizational performance. *Human Resource Development Quarterly*, *22*, 437–458.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, 36*, 717–731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Raidl, M., Johnson, S., Gardiner, K., Denham, M., Spain, K., & Lanting, R. (2004). Use retrospective surveys to obtain complete data sets and measure impact in extension programs. *Journal of Extension*, *42*(2). Retrieved October 19, 2005, from http://www.joe.org/joe/2004april/rb2.shtml

Saks, A. M., & Gruman, J. A. (2014). What do we really know about employee engagement? *Human Resource Development Quarterly*, *25*, 155–182.

Schmidt, F. L., Le, H., & Oh, I.-S. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 317–333). New York: Russell Sage Foundation.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 257–277). New York: Russell Sage Foundation.

Shuck, B., Twyford, D., Reio, T. G., & Shuck, A. (2014). Human resource development practices and employee engagement: Examining the connection with employee turnover intentions. *Human Resource Development Quarterly*, *25*, 239–270.

Sommer, K. L., & Kulkarni, M. (2012). Does constructive performance feedback improve citizenship intentions and job satisfaction? The roles of perceived opportunities for advancement, respect, and mood. *Human Resource Development Quarterly*, *23*, 177–201.

Song, J. H., Kolb, J. A., Lee, U. H., & Kim, H. K. (2012). Role of transformational leadership in effective organizational knowledge creation practices: Mediating effects of employees' work engagement. *Human Resource Development Quarterly*, *23*, 65–101.

Sookhai, F., & Budworth, M. (2010). The trainee in context: Examining the relationship between self-efficacy and transfer climate for transfer of training. *Human Resource Development Quarterly*, *21*, 257–272.

Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem. *Journal of Applied Psychology*, *72*, 438–443.

Thompson, B. (2002a). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage Publications.

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24–32.

Trudel, J., & Reio Jr., T. G. (2011). Managing workplace incivility: The role of conflict management styles-antecedent or antidote? *Human Resource Development Quarterly*, *22*, 395–423.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*, 159–168.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509–522.

Walsh, B. M., Bauerle, T. J., & Magley, V. J. (2013). Individual and contextual inhibitors of sexual harassment training motivation. *Human Resource Development Quarterly*, *24*, 215–237.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. (Reprint available through the APA home page http://www.apa.org/pubs/journals/amp/)

Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: Effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Processes*, *57*, 185–206.

Woolf, S. H. (2008). The meaning of translational research and why it matters. *Journal of the American Medical Association, 29*, 211–213.

Yamkovenko, B., & Holton, E. (2010). Toward theoretical model of dispositional influences on transfer of learning: A test of a structure model. *Human Resource Development Quarterly*, *21*, 381–410.

Kim F. Nimon is Associate Professor in the Department of Human Resource Development and Technology at The University of Texas at Tyler. Her areas of expertise encompass workforce development (human resource development, employee engagement, corporate chaplaincy programs) and analytical methodologies (applied general linear model analyses, statistical software programming, measurement and evaluation). She has published extensively in refereed journals including American Journal of Evaluation, Frontiers in Psychology, Human Resources Quarterly, Multivariate Behavioral Research, *and* Organizational Research Methods. *Dr. Nimon has made over 80 refereed presentations for international and national scholarly organizations and was awarded the Early Career Scholar Award by the Academy of Human Resource Development. She also holds patents from both the U.S. Patent and Trademark Office and Canada Intellectual Property Office. Dr. Nimon has provided leadership in professional organizations including AERA, AHRD, and SERA. She also serves as Associate Editor for* HRDQ.

Marina Astakhova is Assistant Professor of Management at The University of Texas at Tyler. Her primary research explores person-environment fit, work passion and organizational sustainability. Her research has been published in Human Relations, Journal of Business Research, Journal of Business and Psychology, Journal of Business Ethics, Journal of Managerial Psychology *and others. Her articles were published in 2013 and 2015 Best Paper Proceedings of the Academy of Management and received Best Poster award (2014) from the Society for Industrial and Organizational Psychology. She published two book chapters on heavy work investment and organizational sustainability.*

Corresponding author:

Kim F. Nimon can be contacted at knimon@uttyler.edu