Received: 25 May 2018 | Revised: 22 October 2019 | Accepted: 29 October 2019

DOI: 10.1111/peps.12372

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Maybe not so independent after all: The possibility, prevalence, and consequences of violating the independence assumptions in psychometric meta-analysis

Zhenyu Yuan[1] | Frederick P. Morgeson[2] | James M. LeBreton[3]

[1]Department of Managerial Studies, College of Business Administration, University of Illinois at Chicago, Chicago, Illinois

[2]Department of Management, Eli Broad College of Business, Michigan State University, East Lansing, Michigan

[3]Department of Psychology, The Pennsylvania State University, University Park, Pennsylvania

Correspondence
Zhenyu Yuan, Department of Managerial Studies, College of Business Administration, University of Illinois at Chicago, 601 S. Morgan Street, University Hall 2219 (MC 243), Chicago, IL 60607.
Email: zyuan19@uic.edu.

Abstract

Psychometric meta-analysis assumes that moderators are unrelated to study artifacts (e.g., criterion reliability), and that study artifacts are independent of true validities. Meeting these assumptions is important for researchers seeking to accurately partition the variance in effect sizes due to study artifacts from the variance due to meaningful moderators. Despite the critical role of these assumptions, we know very little about their tenability. To address this basic gap in the literature, we conducted three studies to determine if there are potential violations of the independence assumptions (Study 1), the prevalence of such violations (Study 2), and the consequences of violating the independence assumptions via a series of Monte Carlo simulations (Study 3). We found that violations of the independence assumptions are not only plausible but also routinely detected across a wide array of topics. Simulation results indicate that violating the independence assumptions can result in biases under certain circumstances, which are further accentuated due to the lack of stability in the estimators. We offer suggestions for the future use of psychometric meta-analysis and discuss the implications for research focused on refining psychometric meta-analysis.

KEYWORDS

moderator, psychometric meta-analysis, validity generalization

1 | INTRODUCTION

Psychometric meta-analysis involves quantitatively summarizing empirical evidence by correcting observed effect sizes for sampling error and statistical artifacts such as range restriction and unreliability (Schmidt & Hunter, 2015). There is little doubt that psychometric meta-analysis has been one of the most influential innovations in the fields of

management and industrial/organizational (I/O) psychology since it was introduced more than 40 years ago (Schmidt & Hunter, 1977). This influence is manifested in numerous ways. For example, meta-analytic summaries of effect sizes, their heterogeneity, and potential moderators are invariably among the most highly cited articles in a given research area (Judge, Cable, Colbert, & Rynes, 2007; Podsakoff, Podsakoff, Mishra, & Escue, 2018). The psychometric meta-analytic method is predominant, with the vast majority of published meta-analyses in management and I/O psychology utilizing this method (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011).

As with any statistical method, a number of important assumptions underlie its use (Schmidt & Hunter, 2015). For example, in the most recent textbook on psychometric meta-analysis, Schmidt and Hunter (2015, p. 88) noted "the nature of artifacts is such that, in most research domains, the artifact values will be independent across studies." In essence, psychometric meta-analysis assumes that study artifacts are independent across situations (i.e., not related to any moderators) and thus unrelated to true correlations. Despite these assumptions, "the VG [validity generalization][1] mathematical models were introduced without first having been tested for the critical independence assumption on which these residualization-based approaches rest" (James, Demaree, Mulaik, & Ladd, 1992, p. 9). As such, it is important to understand whether the independence assumptions hold, and if not, the implications of violating the independence assumptions.

A small body of research explored this issue over a 10-year period starting in the mid-1980s (Burke, Rupinski, Dunlap, & Davison, 1996; James, Demaree, & Mulaik, 1986, 1992; James, Demaree, Mulaik, & Ladd, 1992; James, Demaree, Mulaik, & Mumford, 1988; Kemery, Mossholder, & Roth, 1987). One simulation study indicated that violation of the independence assumptions results in inaccurate credibility interval estimates (Raju, Anselmi, Goodman, & Thomas, 1998), which play a critical role in accurately interpreting meta-analytic findings (Whitener, 1990). However, there has been little subsequent work in this area (see Köhler, Cortina, Kurtessis, & Gölz, 2015, for an exception), thus precluding a complete understanding of the consequences associated with independence violations.

In a search of meta-analyses published in top I/O psychology and management journals between 2015 and 2018, we found that 79% used psychometric meta-analysis.[2] Although this number attests to the unabated popularity of psychometric meta-analysis, we did not locate a single meta-analysis that discussed or empirically tested the underlying independence assumptions. As such, it appears that its users have failed to test the tenability of this statistical tool's fundamental assumptions.

Taken as a whole, this research gap not only precludes critical evaluations and possible refinement of psychometric meta-analysis, but also hinders scholarly understanding and applications of this important method. In addition, inaccurate meta-analytic estimates and results may have inadvertently found their way into the literature. To address these concerns, we conducted three studies to examine the tenability of independence assumptions and the consequences of violating independence assumptions. In Study 1, we conduct a reliability generalization study to investigate a key assumption underlying psychometric meta-analysis (i.e., that study artifacts such as criterion reliability are not correlated with moderators). After finding violations to this assumption, in Study 2 we explore the prevalence of violations of independence assumptions among published meta-analyses. In Study 3, we conduct a large-scale Monte Carlo simulation study to understand the consequences of violating independence assumptions.

2 | PSYCHOMETRIC META-ANALYSIS AND ITS ASSUMPTIONS

Psychometric meta-analysis assumes that study artifact parameters are independent of true correlations and that study artifact parameters are independent of each other (Pearlman, Schmidt, & Hunter, 1980; Schmidt & Hunter, 2015; Schmidt, Pearlman, & Hunter, 1980). A corollary of these assumptions is that moderators,[3] which may be associated with true correlations, are independent of study artifacts (Raju et al., 1998; Raju, Pappas, & Williams, 1989).

To understand how this process works, take the example of cognitive ability and job performance. Past research has shown that the relationship between general cognitive ability and job performance is stronger in more complex jobs (Hunter, 1986). As such, job complexity acts as a moderator of the cognitive ability to performance relationship. In a

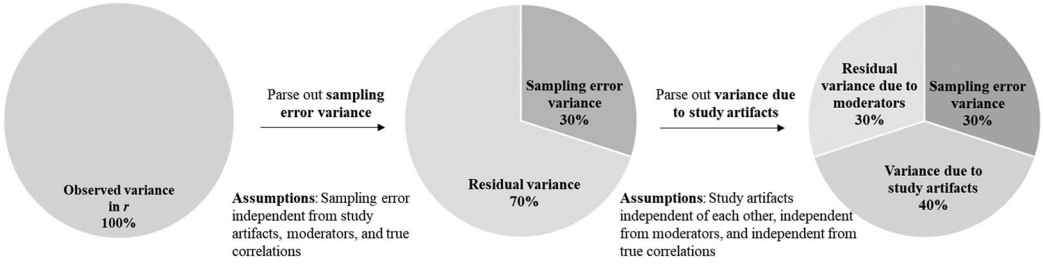

FIGURE 1 Illustration of the variance residualization process of psychometric meta-analysis
Note. Percentages of sampling error variance and variance attributed to study artifacts are for illustrative purposes only.

typical psychometric meta-analysis, researchers would conduct the following analyses (see Figure 1). First, they would calculate (a) the variance in observed effect sizes (i.e., *r*), (b) the variance due to sampling error, and (c) the variance due to study artifacts. Second, they would compute the residual variance in *r* (i.e., the variance in *r* that remains after subtracting the variance due to sampling error and study artifacts). Third, they would attribute this residual variance to moderators (i.e., true effect size heterogeneity). Finally, researchers could take different analytical approaches to identify whether job complexity is a plausible moderator (Cortina, 2003).

Importantly, in this process, it is assumed job complexity is the "only" factor responsible for true effect size heterogeneity and that study artifacts (e.g., criterion reliability) are not systematically different in jobs that are more (vs. less) complex (see Figure 2a). This assumption is essential when correcting for study artifacts because if it is not true, these variance components are confounded. As a result, researchers will have incorrectly attributed true variation due to the moderator to error variation associated with various statistical artifacts (see Figure 2b). Critically, there is no guarantee that independence conditions will always be met; however, psychometric meta-analysis simply assumes so. Nevertheless, the violation of independence assumptions can occur for many reasons. For example, criterion reliability may covary with job complexity such that the criterion space for less complex jobs (e.g., package delivery) may be easier to identify and reliably measure (e.g., number of packages correctly delivered on time). In contrast, more cognitively complex jobs (e.g., financial analyst) may have more complicated criterion spaces that are comparatively more difficult to measure in a reliable manner (e.g., measured through convenient yet unreliable measurement systems such as supervisory ratings).

Challenging the assumption that study artifacts are uncorrelated with moderators, James, Demaree, Mulaik, and Ladd (1992) provided a theoretical account of how a moderator (i.e., restrictiveness of climate) might constrain the expression of individual differences. This would result in smaller criterion variance and lead to lower criterion reliability estimates (in addition to affecting the true correlation). Empirical evaluation of this model found mixed evidence for the independence of study artifacts and moderators, calling for additional research (Burke et al., 1996). Despite this call, the role of the independence assumptions has received little empirical attention. We seek to further empirically evaluate the assumption that moderators are unrelated to study artifacts. Although the independence assumption applies to all study artifacts, our review of the literature suggests criterion reliability is a particularly important study artifact to investigate (Burke et al., 1996; James, Demaree, Mulaik, & Ladd, 1992). Therefore, we focus on criterion reliability and explore its relationship to moderators in Study 1.

3 | STUDY 1[4]

We conducted a reliability generalization study on safety climate (i.e., the moderator) and safety performance reliability. The use of reliability generalization is necessary because we need to accumulate criterion reliability and relate it to the moderator across different samples. Reliability generalization recognizes that coefficient alphas from different

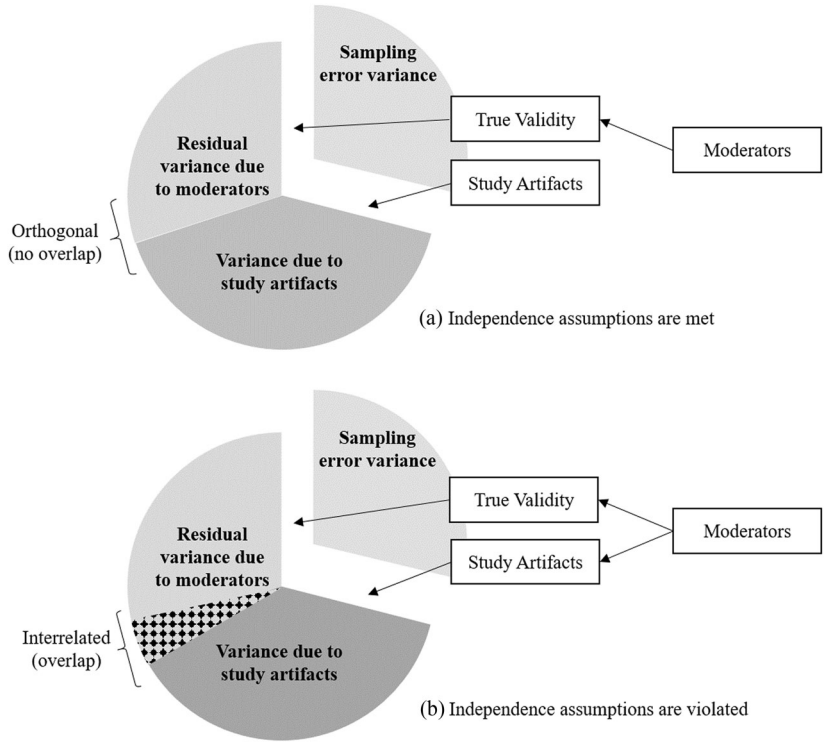

FIGURE 2 Illustration of violating the independence assumptions in the variance residualization process
Note. Partly based on Figure 1 in "Validity generalization in the context of situational models," by L. R. James, R. G. Demaree, S. A. Mulaik, and R. T. Ladd, 1992, *Journal of Applied Psychology*, 77, p. 9. Copyright 1992 by the American Psychological Association. Adapted with permission.

studies should be weighted by its precision, the omission of which would result in biased estimates (Greco, O'Boyle, Cockburn, & Yuan, 2018; Rodriguez & Maeda, 2006). We examined safety climate and safety performance because of our experience in this research area. Several features of this literature also make this area suitable for an empirical evaluation of the independence assumption: (a) there exists a widely agreed-upon conceptualization of safety performance in safety research (Griffin & Neal, 2000; Neal, Griffin, & Hart, 2000); (b) there is a sufficiently large empirical base of studies on both safety climate and safety performance; and (c) the sample mean of safety climate provides a straightforward way to operationalize the overall moderator for a given sample.

Safety climate is an important aspect of organizational climate that is focused on safety (Zohar, 1980, 2000) and is an active research area (Jiang, Lavaysse, & Probst, 2019; Schneider, González-Romá, Ostroff, & West, 2017). Safety climate can be defined as "a summary of molar perceptions that employees share" about the safety-relevant aspects of their work environments (Zohar, 1980, p. 96). Thus, safety climate encompasses shared perceptions of various aspects of the work environment regarding safety procedures such as safety training and required work pace (Zohar, 1980). Greater levels of safety climate indicate a work environment with a stronger emphasis on workplace safety. It has been shown to be an important situational variable moderating the relationship between various predictors and safety behaviors. For example, safety climate strengthens the within-unit relationship between leader–member exchange and safety citizenship role definitions (Hofmann, Morgeson, & Gerras, 2003) and the within-unit relationship between coworkers' normative influence and individuals' own safety performance (Jiang, Yu, Li, & Li, 2010).

The criterion (i.e., safety performance) taps into employee work behaviors that are important for the maintenance and promotion of a safe working environment. Griffin and Neal (2000; see also Andriessen, 1978; Marchand, Simard, Carpentier-Roy, & Ouellet, 1998) conceptualized safety performance as a two-dimension construct, based in part

on the well-established distinctions in the job performance literature between task performance and contextual performance (Borman & Motowidlo, 1993). *Safety compliance* refers to core activities that individuals engage in to maintain workplace safety (similar to task performance). Examples include abiding by safety procedures and wearing personal protective equipment. In contrast, *safety participation* taps into the voluntary citizenship behaviors that help to promote workplace safety. Examples include putting in extra effort to promote safety and voluntarily carrying out tasks that improve workplace safety. We incorporate both safety compliance and safety participation as criterion variables in Study 1.

3.1 | Method

3.1.1 | Literature search

Relevant studies were first identified by searching PsycINFO, Google Scholar, and ProQuest Dissertations and Theses up to December 2017. The following keywords were used: "safety," "safety performance," and "safety behavior." We also conducted an exhaustive legacy search (i.e., studies that cited the original scale development papers) of Griffin and Neal (2000), Neal et al. (2000), and Neal and Griffin (2006) as the safety performance scale was initially developed and used in these published papers. The reference lists of past meta-analytic reviews on safety performance (Beus, Dhanani, & McCord, 2015; Christian, Bradley, Wallace, & Burke, 2009; Nahrgang, Morgeson, & Hofmann, 2011) were closely examined to identify articles that were not included in the initial search. We also made requests via listservs to members of the Human Resources Division and Organizational Behavior Division of Academy of Management and the Occupational Health Psychology network for unpublished studies and datasets. Given that their conceptualization of safety performance is the most influential and commonly used in this literature, we elected to only include studies that relied on the safety performance scale of Griffin and Neal (2000; Neal et al., 2000; Neal & Griffin, 2006) to test both safety compliance and safety participation as criterion variables. Doing so helps us to minimize the problems associated with divergent operationalizations of safety behaviors.

3.1.2 | Inclusion criteria and coding of studies

To be included, an empirical study would need to report the mean of safety climate. Furthermore, information about the number of scale anchors (e.g., 5-point or 7-point scale) and the numerical value of the starting point must be reported for a study to be included. Further, a study must also report the reliability of the safety performance dimensions (i.e., information regarding criterion reliability must be available) and the number of scale items—both of which are required for reliability generalization estimation. Using these criteria, we identified 39 independent samples for safety compliance and 37 for safety participation (see Supporting Information). We coded for sample size, the means and scale information (i.e., starting value and range) of safety climate, the internal consistency reliability of safety compliance and participation (i.e., coefficient alpha), and the number of items. In cases where the overall mean of safety climate was not available and only the descriptive statistics of multiple dimensions of safety climate were reported, we coded the mean of management commitment to safety—the core dimension of safety climate (Zohar, 2002).

3.1.3 | Analytic procedure

Prior to analysis, we recoded the sample mean of safety climate on a zero to one metric so that the values from studies using different scales (e.g., 5-point vs. 7-point) were comparable. This value provides an operationalization of the moderator for the respective study sample, similar to previous research that used the overall mean from each unit to capture unit-level climate (Burke et al., 1996). To formally test the effect of safety climate on criterion reliability, we conducted meta-analytic regression with a random effect, restricted maximum likelihood estimator. We did this because simulation evidence suggests it provides a good balance between unbiasedness and efficiency in detecting effect size heterogeneity (Thompson & Sharp, 1999; Viechtbauer, 2005). We used the built-in command for reliability generalization meta-analysis (Vacha-Haase, 1998) in the open-source R package "metafor" developed by Viechtbauer

TABLE 1 Meta-analysis of safety climate and safety performance reliability (Study 1)

Descriptive statistics				
	Safety climate	Safety compliance reliability	Safety climate	Safety participation reliability
Mean	.68	.85	.68	.80
SD	.11	.11	.10	.08
k	39		37	
N	19,089		17,545	
Pearson's *r*	$r = .34$ ($t = 2.20$, $df = 37$, $p = .03$)		$r = .18$ ($t = 1.11$, $df = 35$, $p = .27$)	

Meta-analytic regression results				
	Model 1a ($k = 39$, $N = 19,089$)		Model 2a ($k = 37$, $N = 17,545$)	
	B	*SE*	*B*	*SE*
Intercept	.62**	.10	.70**	.08
Safety climate	.34*	.15	.16	.12
	Model 1b ($k = 37$, $N = 17,892$)		Model 2b ($k = 35$, $N = 16,348$)	
	B	*SE*	*B*	*SE*
Intercept	.57**	.12	.65**	.11
Safety climate	.41*	.17	.22	.15

Note. *SD* = standard deviation; *k* = number of independent samples; *N* = total sample size; *df* = degrees of freedom. Descriptive statistics (Mean, *SD*, and Pearson's *r*) were based on coded sample-level information (i.e., not weighted by sample size). Random effect meta-analytic regression was based on restricted maximum-likelihood estimator. Unstandardized coefficients are reported. Models 1a and 2a were based on the full set of samples, whereas influential samples were excluded when estimating Models 1b and 2b.
*$p < .05$; **$p < .01$.

(2010). In conducting reliability generalization meta-analytic regression, we screened for possible outliers using the "influence.rma.uni" command in this package (see Viechtbauer, 2010 for details). Upon identifying outliers, we examined the reason that a study was identified as an influential case by examining the various diagnostic statistics. We then conducted the meta-analytic regression, and presented findings both including and excluding these cases to assess their impact on the study findings (Aguinis, Gottfredson, & Joo, 2013; Cortina, 2003).

3.2 | Results

The descriptive statistics of sample safety climate and safety performance reliability coded from each included sample are reported in the upper half of Table 1. As a preliminary examination of the trend of covariation between safety climate and criterion reliability, we estimated the Pearson's correlation. The raw Pearson's correlation estimates showed substantial interdependence between sample safety climate and criterion reliability for safety compliance ($r = .34$, $p < .05$), whereas the relationship between sample safety climate and the reliability of safety participation was positive but not significant ($r = .18$, $p = .27$).

3.2.1 | Safety compliance

Meta-analytic regression was used to test the effect of safety climate on criterion reliability. The effect of safety climate on the reliability of safety compliance was positive and statistically significant ($B = .34$, $SE = .15$, $p < .05$; Model 1a in Table 1). Two studies were identified as influential cases, with the leave-one-out diagnostics suggesting that they were both prediction outliers (Aguinis et al., 2013).[5] To provide a more conservative estimate and check for the robustness of findings based on the full set of studies, we excluded these two samples and reran the regression. The effect of safety climate was still statistically significant ($B = .41$, $SE = .17$, $p < .05$; Model 1b in Table 1) with these samples excluded.

3.2.2 | Safety participation

The effect of safety climate on the reliability of safety participation was not statistically significant ($B = .16$, $SE = .12$, $p = .20$; Model 2a in Table 1). The outlier check showed that the same two studies identified as outliers in the previous analysis were again influential cases, with the leave-one-out diagnostics suggesting that they were prediction outliers. Accordingly, we reran the regression excluding these samples. The effect of safety climate was still not statistically significant ($B = .22$, $SE = .15$, $p = .14$; Model 2b in Table 1) after excluding these samples.

3.3 | Discussion

Overall, results regarding safety compliance as the criterion suggest that the relationship between the moderator (i.e., safety climate) and an important study artifact (i.e., safety compliance reliability) was not statistically independent. Thus, findings from Study 1 suggest a basic assumption underlying psychometric meta-analysis does not always hold, at least in the context of safety climate and safety compliance. However, it is unclear the extent to which violation of independence assumptions is more broadly prevalent in the published literature. To explore this issue, in Study 2 we examine recently published psychometric meta-analyses.

4 | STUDY 2[6]

The goal of Study 2 is to survey recently published psychometric meta-analyses and assess the extent to which the independence assumptions may have been violated. To broaden our focus, we examine the independence among moderators, predictor reliability, and criterion reliability because both study artifacts are commonly corrected for (Köhler et al., 2015).

4.1 | Method

We contacted authors who published a psychometric meta-analysis from 2015 to February 2018 in four leading management and I/O journals (*Academy of Management Journal*, *Journal of Applied Psychology*, *Journal of Management*, and *Personnel Psychology*). We chose this time frame and these journals to provide an illustrative assessment of the potential violation of the independence assumptions. Studies that did not correct for unreliability based on information from individual studies were excluded, as hypothetical artifact distributions provide little information about the empirical distribution of study artifacts. We asked each author to focus on the effect size that had the largest number of independent samples in the meta-analysis, and run the correlations among predictor reliability, criterion reliability, and any moderators that were incorporated. In other words, the correlation analysis was conducted at the between-sample level. Each independent sample was an observation with variables including predictor reliability, criterion reliability, and moderators (if applicable).

We sent an email detailing our request for help with additional analyses to the corresponding author of each meta-analysis, followed by a reminder sent out 2 weeks later. Further, in accordance with the *Journal of Applied Psychology's* reporting guidelines, the entire coding sheet is available for most meta-analyses published in this journal. Therefore, in cases where we did not hear back from the corresponding author of *Journal of Applied Psychology* articles, we used the supplementary material for our analysis. We were able to obtain the results from 17 meta-analyses, out of 26 that were deemed relevant for our purpose (response rate = 65.38%).

4.2 | Results and discussion

Results are presented in Figures 3 and 4. In each figure, the *x*-axis refers to the number of independent samples in the meta-analysis for the correlation coefficient, whereas the *y*-axis indicates the magnitude of the correlation. In Figure 3, each dot refers to a correlation between predictor/criterion reliability and a moderator. In most cases, there is more

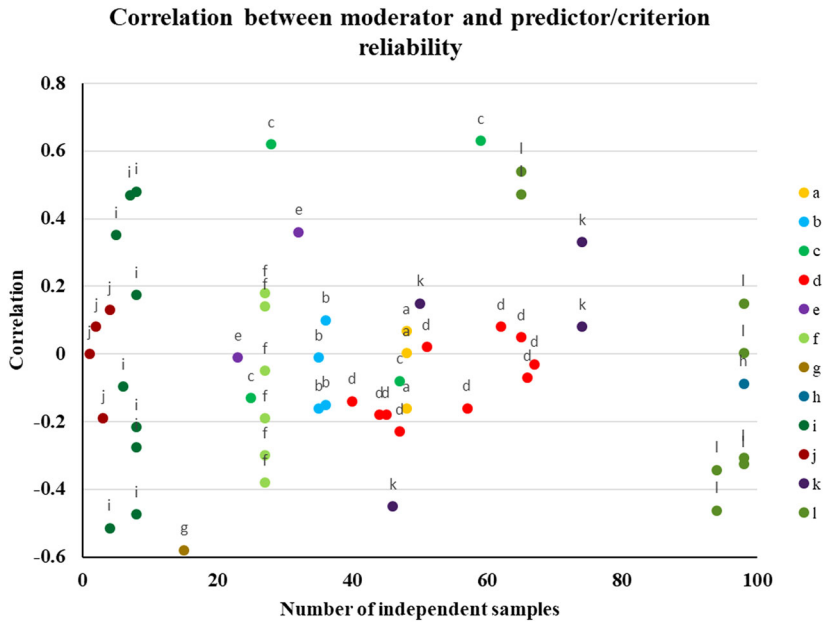

FIGURE 3 Correlation between moderator and predictor/criterion reliability from published meta-analyses (Study 2)

Note. Each dot corresponds to a correlation between a moderator and predictor/criterion reliability; letters "a" to "l" refer to different published meta-analyses; the letter above each dot indicates the source of each correlation; the correlation is plotted along the *y*-axis, whereas the *x*-axis indicates the number of independent samples for the correlation. There may be multiple correlations (i.e., multiple dots) from the same meta-analysis as there may be more than one moderator incorporated in the meta-analysis and/or both predictor and criterion reliability estimates are available from the meta-analysis.

than one correlation from a psychometric meta-analysis, as both predictor and criterion reliability were reported or more than one moderator was incorporated. Thus, correlations from the same meta-analysis are represented with the same color. In Figure 4, each dot refers to the correlation between predictor reliability and criterion reliability, with each meta-analysis contributing only one correlation.

Deviation from 0 on the *y*-axis in either direction suggests potential violation of the independence assumptions as the correlations between moderators and predictor/criterion reliability (Figure 3) and between predictor and criterion reliability (Figure 4) were nonzero. If the independence assumption always holds, one would expect to see the data points distributed relatively close to the baseline where $y = 0$. However, this was not the case in either of the two figures. In fact, the distribution of the dots shown in Figures 3 and 4 indicates the relationships among study artifacts and moderators were nontrivial in many cases. It might be tempting to view the very strong correlations as outliers and thus as unusual observations that could be eliminated or discounted in some way. Yet, this would be inadvisable because each dot reflects estimated covariation among moderators, predictor reliability, and criterion reliability at the population level from a published meta-analysis. Thus, their existence provides direct support that there can be substantial violations of the independence assumptions.

Next, to assess whether the violations were unique to a specific research area, we coded the predictors, criterion variables, and moderators (see Supporting Information for complete results). Consistent with James, Demaree, Mulaik, and Ladd's (1992) general concern about the tenability of independence assumptions, moderators that demonstrated a lack of independence from predictor/criterion reliability included contextual factors at different levels (e.g., culture; team design features; occupational characteristics) and methodological factors (e.g., rating purpose of performance appraisals). As such, Study 2 highlights that violating the independence assumptions is not solely limited to

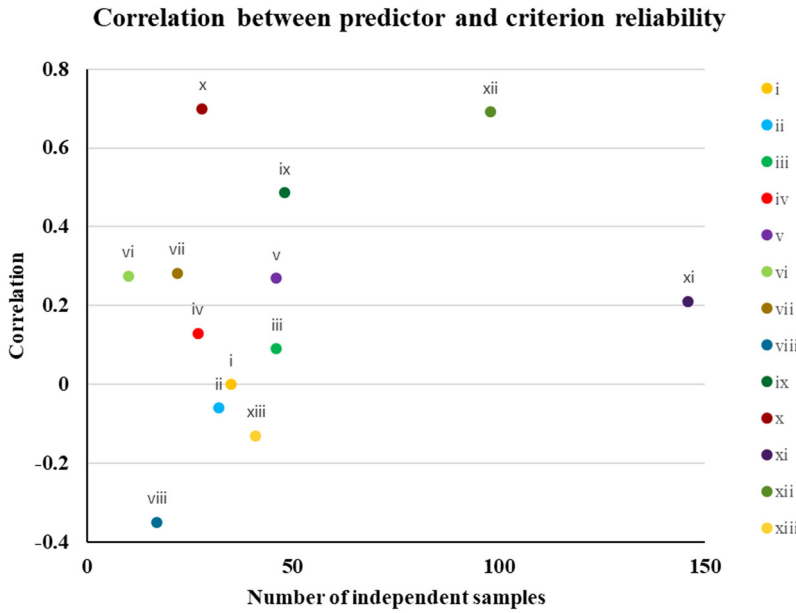

FIGURE 4 Correlation between predictor reliability and criterion reliability from published meta-analyses (Study 2)

Note. Each dot corresponds to a correlation between predictor reliability and criterion reliability; letters "i" to "xiii" refer to different published meta-analyses; the letter above each dot indicates the source of each correlation; the correlation is plotted along the *y*-axis, whereas the *x*-axis indicates the number of independent samples for the correlation. Each meta-analysis contributes only one correlation (i.e., one dot) between predictor reliability and criterion reliability.

a specific research area, but rather appears to be more widespread than previously assumed. Similarly, the predictors were not limited to a specific category as they ranged from personality (e.g., conscientiousness) to leadership (e.g., leader–member exchange) whereby the criterion variables included strain (e.g., general strain), job attitudes (e.g., job satisfaction), and work behaviors (e.g., task performance).

Because moderators may be systematically related to population correlations, a violation of the independence assumption between moderators and criterion reliability will further result in a lack of independence between true validity and study artifacts. At a conceptual level, this violation would suggest that making corrections for attenuation due to unreliability—a critical step in psychometric meta-analysis—is problematic because the variance in effect sizes that is being attributed to criterion unreliability is no longer solely artifactual (James, Demaree, Mulaik, & Ladd, 1992; Spearman, 1910). Instead, some of that variance is meaningful (i.e., true) variance engendered by the moderator variable (see Figure 2b). According to James, Demaree, Mulaik, and Ladd (1992), when independence assumptions are violated, "*it is neither meaningful nor possible to use the VG statistical residualization process to partition variance in the* r_k *into variance due to statistical artifacts and variance due to situational moderators*" (p. 9; emphasis in original). From a pragmatic standpoint, much less is known regarding the extent to which violating the independence assumptions will yield problematic results using psychometric meta-analysis. Therefore, in Study 3 we conducted a large-scale Monte Carlo simulation to explore this issue.

5 | STUDY 3

Preliminary simulation research provides some evidence of biases in psychometric meta-analytic estimates when the independence assumptions are violated. For example, Raju et al. (1998) conducted a simulation study to examine what

TABLE 2 Summary of manipulated factors (Study 3)

Factor	Description	Values
1	Mean true validity ($\bar{\rho}_{xy}$)	.10, .30, .50
2	True validity *SD* (σ_{ρ})	.15, .30
3	Mean criterion reliability (ρ_{yy})	.50, .70, .90
4	Number of independent samples (*k*)	10, 30, 50, 100
5	Correlation between moderator and criterion reliability ($\rho_{mod,\ \rho yy}$)	−.50, −.30, .30, .50
6	Correlation between moderator and true validity ($\rho_{mod,\ \rho xy}$)	−.50, −.30, .30, .50
7	Correlation between criterion reliability and true validity ($\rho_{\rho yy,\ \rho xy}$)	−.50, −.30, .00, .30, .50

Note. These factors created a constellation of $(3 \times 2 - 1) \times 3 \times 4 \times 4 \times 4 \times 5 = 4{,}800$ conditions, as the conditions where $\bar{\rho}_{xy} = .50$ and $\sigma_{\rho} = .30$ were considered unrealistic and thus excluded from the simulation. Within each condition, we sampled 1,000 population matrices.

happens when study artifacts are correlated with true validity. Their findings indicated that violation of the independence assumption resulted in overestimation of true effect size heterogeneity in nine out of 12 simulated conditions. Further, some of the estimation biases were substantial.

However, their study should be considered as only a preliminary step for three reasons. First, they examined the accuracy of the Taylor series approximation 1 (TSA1; Raju & Burke, 1983) procedure, but not psychometric meta-analysis. This was done due to their conclusion that these two procedures tend to yield highly similar estimates (Mendoza & Reinhardt, 1991; Raju & Burke, 1983). Second, TSA1 involves corrections based on artifact distributions, not individual corrections. These procedural differences have thus created some uncertainty regarding the applicability of their simulation findings to psychometric meta-analysis based on individual corrections. Third, they focused on a limited set of conditions, as they did not manipulate mean true validity or the distribution of study artifacts. This makes it difficult to extrapolate their findings across different combinations of true validity and study artifacts.

Therefore, in Study 3 we conducted a series of simulations to examine a much broader set of factors than previously considered. This enables a more complete understanding of the implications of violating the independence assumptions. We also evaluate psychometric meta-analysis based on individual study corrections, as this approach is far more common in recently published meta-analyses than corrections based on artifact distributions.

5.1 | Method

The accuracy of psychometric meta-analytic estimates was examined using Monte Carlo procedures. Specifically, we manipulated seven factors to create a constellation of different experimental conditions, each of which contained the true population values for a given condition. From each condition, we randomly drew one set of values of the seven factors from the population distribution, created the respective number of independent samples, conducted a psychometric meta-analysis using those samples, and then repeated this process 1,000 times for each condition. In choosing the levels of the manipulated factors, we referred to past meta-analyses and aligned our decisions with the published literature. Further, we drew on meta-analytic evidence from both employee selection research and other research areas, as the use of psychometric meta-analysis has become widespread in I/O psychology and management. In doing so, we sought to expand the relevance of our findings to an array of topic domains.

5.1.1 | Manipulated factors

As summarized in Table 2, we manipulated seven factors including: mean true validity ($\bar{\rho}_{xy}$), true validity standard deviation (*SD*; σ_{ρ}), mean criterion reliability (ρ_{yy}), the number of independent samples (*k*), the correlation between moderator and criterion reliability ($\rho_{mod,\ \rho yy}$), the correlation between moderator and true validity ($\rho_{mod,\ \rho xy}$), and the correlation between criterion reliability and true validity ($\rho_{\rho yy,\ \rho xy}$).

To approximate different levels of mean true validity ($\bar{\rho}_{xy}$), reported in the literature, we chose values of .10, .30, and .50, which are similar to the estimated mean true validity of important predictors such as extraversion ($\bar{\rho}_{xy} = .13$ in Barrick & Mount, 1991), job satisfaction ($\bar{\rho}_{xy} = .30$ in Judge, Thoresen, Bono, & Patton, 2001), and general cognitive ability ($\bar{\rho}_{xy} = .51$ for medium-complexity jobs in Hunter & Hunter, 1984), respectively. Following Hall and Brannick (2002), we specified two values for true validity *SD* (σ_ρ) at .15 and .30 to represent situations with moderate (e.g., peer ratings predicting supervisor performance ratings; Hunter & Hunter, 1984) and substantial heterogeneity (e.g., team conflict and team functioning; De Wit, Greer, & Jehn, 2012), respectively. Regarding criterion reliability (ρ_{yy}), we referenced the literature on reliability generalization and specified three distributions (1: $\mu = .50$, $\sigma = .08$; 2: $\mu = .70$, $\sigma = .08$; 3: $\mu = .90$, $\sigma = .08$). This approximates the distributions of interrater reliability of supervisor ratings of job performance (Viswesvaran, Ones, & Schmidt, 1996), the reliability estimates considered minimally acceptable for exploratory research (Nunnally, 1978), and the internal consistency reliability of other-rated job performance (Greco et al., 2018), respectively. Similar to Hall and Brannick (2002), the number of independent samples (*k*) was set at four values (10, 30, 50, and 100). In sum, these four manipulated factors are intended to simulate situations that are likely to be encountered by researchers in I/O psychology and management.

The correlations between moderator and criterion reliability ($\rho_{mod,\ \rho yy}$) and between moderator and true validity ($\rho_{mod,\ \rho xy}$) were set at −.50, −.30, .30, and .50, which covered medium and large correlations (cf., Bosco, Aguinis, Singh, Field, & Pierce, 2015; Cohen, 1962). The correlation between criterion reliability and true validity ($\rho_{\rho yy,\ \rho xy}$) was set at five values (−.50, −.30, .00, .30, and .50). The inclusion of .00 represents the independence assumption (i.e., no relationship between criterion reliability and true validity) and enables a direct comparison between when the assumption is met and when it is violated. These three manipulated factors represent situations where one or more independence assumptions are violated.

As noted by Hall and Brannick (2002), the combination of $\bar{\rho}_{xy} = .50$ and $\sigma_\rho = .30$ represents a highly unrealistic situation, as true correlations, according to this distribution, would be frequently over .80. Such correlations are extremely rare in I/O psychology and management. Further, the resulting distribution of population validity is likely nonnormal. Therefore, this combination was excluded from the simulation conditions, which resulted in a $(3 \times 2 - 1) \times 3 \times 4 \times 4 \times 4 \times 5$ factorial design. This created a constellation of 4,800 conditions.

5.1.2 | Prespecified factors

We specified the distribution of sample size ($\mu = 200$, $\sigma = 25$) to engender sampling error, which is consistent with typical sample sizes in the I/O psychology literature (Bosco et al., 2015). As the measurement unit of the moderator (e.g., 5-point or 7-point scale) was not of primary interest to our simulation, we used a standard normal distribution ($\mu = 0$, $\sigma = 1$). Because the effect of sampling error is assumed to be unsystematic (i.e., it is orthogonal to other study artifacts; Schmidt & Hunter, 2015), we set the population-level relationship between sample size and other factors (i.e., true validity, criterion reliability, and moderator) to zero.

5.1.3 | Monte Carlo procedure

The 4,800 sets of values for the manipulated factors were used in conjunction with the prespecified parameters to create multivariate normal distributions. That is, in each of the 4,800 conditions we first created a multivariate matrix, which served as the population parameters for that condition. This matrix contained the mean, *SD*, and intercorrelations among true validity, criterion reliability, moderator, and sample size. Depending on the number of independent samples (*k*) in that condition, we then drew *k* sets of values from this multivariate matrix.[7] Each of the *k* sets thus contains true validity, criterion reliability, moderator, and actual sample size for that sample. We then followed the procedure described in Law, Schmidt, and Hunter (1994) to create the predictor and criterion scores in that sample, using the set of sample values. Next, we calculated the observed correlation (*r*) from that sample. We compiled the various pieces of information from each sample (observed correlation, actual sample size, and actual criterion reliability) and conducted a psychometric meta-analysis based on the *k* samples, using the "metafor" package developed

by Viechtbauer (2010).[8] Within each condition (i.e., each multivariate matrix), we repeated this process 1,000 times and accumulated 1,000 estimates of $\bar{\rho}_{xy}$ and σ_ρ. Finally, we calculated the average estimated true validity ($\widehat{\overline{\rho_{xy}}}$) and its estimated *SD* ($\widehat{\sigma_\rho}$) within each condition.

5.1.4 | Bias and stability

We first consider bias, defined as the difference between the population parameters and the estimated sample statistics obtained from our simulations (i.e., $\widehat{\overline{\rho_{xy}}} - \bar{\rho}_{xy}$ and $\widehat{\sigma_\rho} - \sigma_\rho$). Bias represents the degree (on average) to which sample estimates deviate from the true population values (Oswald & Johnson, 1998). A positive score suggests overestimation (i.e., an upward bias), whereas a negative score indicates underestimation (i.e., a downward bias). Additionally, we investigate stability, which taps into the extent to which the estimates from a specific meta-analysis may deviate from the average bias. In other words, bias captures the central tendency of psychometric meta-analytic estimates across many iterations of meta-analyses, whereby stability can be understood as a measure of variability of estimates (Oswald & Johnson, 1998). Both bias and stability furnish a complete understanding of the practical implications of violating the independence assumptions, in that the average degree of bias should not be equated with the bias that will manifest in a given meta-analysis (Gillespie, Oswald, & Converse, 2002; Oswald & Johnson, 1998).

5.2 | Analysis and results

We first identified the factors that could influence estimation bias and then turned to the analysis of stability. Next, we considered both bias and stability to interpret the practical implications of independence violations.

5.2.1 | Bias

We conducted a set of ordinary least square (OLS) regression analyses to examine which manipulated factors could explain variance of the biases of the estimated mean true validity ($\widehat{\overline{\rho_{xy}}} - \bar{\rho}_{xy}$) and its *SD* ($\widehat{\sigma_\rho} - \sigma_\rho$). We also included the two-way interactions between factors. Each manipulated condition consisted of an observation, with the seven manipulated factors and their interactions as predictors, and the estimation bias of mean true validity and its *SD* as the two outcome variables. The total sample size for the OLS regression analyses was 4,800.

Regarding the estimation bias of mean true validity ($\widehat{\overline{\rho_{xy}}} - \bar{\rho}_{xy}$), we first entered the main effects of the seven manipulated factors (Model 1a in Table 3). Next, we explored two-way interactions between different factors. As there were many possible two-way interactions to be included, we ran all analyses first and only retained significant main and moderation effects in Model 1b. Due to the large sample size, we had sufficient statistical power to detect any relevant predictors. In other words, the nonsignificant predictors did not have any pronounced influence and thus could be dropped from the analyses. We recognize that this approach is capitalizing on sampling error, but given the number of possible higher-order interactions, we balanced this concern against a desire for a more parsimonious model. Adding the interactions in Model 1b explained an additional 68% of the total variance. Model 1b results indicated that mean true validity, its *SD*, mean criterion reliability, the correlation between criterion reliability and true validity, and their interactions all played a role in influencing the estimation bias of mean true validity. Notably, the correlation between criterion reliability and true validity, which directly speaks to the independence assumption between these two components, was a significant predictor in Model 1a, and also interacted with mean criterion reliability and true validity *SD* in Model 1b. Thus, violating the independence assumption was associated with estimation bias of mean true validity and this relationship also depended on the levels of mean criterion reliability and true validity *SD*.

Regarding the estimation bias of true validity *SD* ($\widehat{\sigma_\rho} - \sigma_\rho$), Model 2a includes the main effects of the manipulated factors. In Model 2b, we retained significant main effects and two-way interactions between manipulated factors. Compared with Model 2a, Model 2b explained an additional 8% of the variance. Specifically, factors including mean true validity, its *SD*, mean criterion reliability, the number of independent samples, the correlation between criterion reliability and true validity, and their interactions were significant. Although in Model 2a the correlation between criterion

TABLE 3 Ordinary least square (OLS) regression results predicting estimation bias (Study 3)

	Estimation bias of mean true validity			
	Model 1a		Model 1b	
	B	*SE*	*B*	*SE*
Intercept	.0063**	.0009	−.0027**	.0006
Main effects				
$\bar{\rho}_{xy}$	−.0085**	.0011	.0299**	.0017
σ_{ρ}	−.0327**	.0022	.0142**	.0022
ρ_{yy}	.0008	.0009	.0008	.0005
k	.0000	.0000		
$\rho_{mod,\ \rho yy}$	.0000	.0004		
$\rho_{mod,\ \rho xy}$	.0001	.0004		
$\rho_{\rho yy,\ \rho xy}$	.0082**	.0004	.1059**	.0011
Interaction terms:				
$\rho_{\rho yy,\ \rho xy} \times \sigma_{\rho}$			.0382**	.0027
$\rho_{\rho yy,\ \rho xy} \times \rho_{yy}$			−.1510**	.0012
$\bar{\rho}_{xy} \times \sigma_{\rho}$			−.2134**	.0088
R^2	.11**		.79**	
ΔR^2			.68**	
	Estimation bias of true validity *SD*			
	Model 2a		Model 2b	
	B	*SE*	*B*	*SE*
Intercept	.0071**	.0004	−.0060**	.0009
Main effects:				
$\bar{\rho}_{xy}$	−.0057**	.0004	.0352**	.0012
σ_{ρ}	−.0732**	.0009	−.0114**	.0037
ρ_{yy}	−.0081**	.0004	.0022*	.0011
k	.0001**	.0000	.0000**	.0000
$\rho_{mod,\ \rho yy}$	.0000	.0002		
$\rho_{mod,\ \rho xy}$	−.0001	.0002		
$\rho_{\rho yy,\ \rho xy}$	−.0002	.0002	−.0031**	.0006
Interaction terms:				
$\bar{\rho}_{xy} \times \sigma_{\rho}$			−.2273**	.0064
$\rho_{yy} \times \sigma_{\rho}$			−.0423**	.0045
$\rho_{\rho yy,\ \rho xy} \times \rho_{yy}$			.0041**	.0009
$k \times \sigma_{\rho}$			.0004**	.0000
$k \times \rho_{yy}$			.0000**	.0000
R^2	.68**		.77**	
ΔR^2			.08**	

Note. N = 4,800. Unstandardized coefficients reported. *SE* = standard error; $\bar{\rho}_{xy}$ = mean true validity; σ_{ρ} = true validity *SD*; ρ_{yy} = mean criterion reliability; k = number of independent samples; $\rho_{mod,\ \rho yy}$ = correlation between moderator and criterion reliability; $\rho_{mod,\ \rho xy}$ = correlation between moderator and true validity; $\rho_{\rho yy,\ \rho xy}$ = correlation between criterion reliability and true validity.

*$p < .05$; **$p < .01$.

reliability and true validity was not a significant predictor, it interacted with mean criterion reliability in Model 2b, suggesting that violating this independence assumption was especially problematic at certain levels of criterion reliability.

Because the nonsignificant factors did not substantially influence estimation bias, results from different levels of these nonsignificant factors can be collapsed for the ease of interpretation. For estimated mean true validity, we collapsed the 4,800 conditions across the nonsignificant factors and obtained 75 mean estimation bias values. These pieces of information are presented in the 75 numbered conditions in the Supporting Information. Similarly, we summarized the estimates of true validity *SD* across the categories of nonsignificant factors and obtained 300 mean estimation bias values. Compared with the factors that influenced the bias of estimated mean true validity, the number of independent samples was also a pertinent factor affecting the estimation bias of true validity *SD*. Accordingly, we presented relevant results within each of the 75 numbered conditions, as a function of the number of independent samples (see Supporting Information).

5.2.2 | Stability

One pertinent factor influencing the variability of estimates is the number of independent samples. Similar to the notion from the central limit theorem that increasing the sample size reduces the variability of its distribution, having more independent samples can also increase stability, with everything else held equal (Oswald & Johnson, 1998). Therefore, we incorporated this factor and used the same four levels (i.e., 10, 30, 50, and 100) to understand its role. Specifically, the number of independent samples, along with other factors identified as relevant in influencing bias, created the different simulated conditions to examine stability. We followed the same Monte Carlo procedure except that we repeated the process 100 times—that is, we conducted 100 iterations for each simulated condition. We did so to provide a convenient way to illustrate how many out of 100 meta-analyses would produce an estimate that fell within a certain range from the true value, although we note that this number of iterations is still sufficiently large to produce trustworthy results in Monte Carlo simulations.

Following Oswald and Johnson (1998), we used the *SD* of the estimates across 100 repetitions to capture stability, with greater *SD* values reflecting less stable estimates. Further, we also calculated the frequency with which estimated mean true validity and its *SD* fell outside the ±.05 intervals of the true population values. This statistic provides a straightforward illustration of how a lack of stability may influence the actual estimates from the 100 iterations. Ideally, one would hope 0% of the iterations will yield an estimate outside the ±.05 intervals of the true population values, as the choice of ±.05—though arbitrary—should be considered a rather liberal standard. In other words, higher percentages represent less stable estimates, as actual estimates frequently deviate from the true values by at least .05. Complete results regarding stability are reported in the Supporting Information within each of the 75 numbered conditions, under each level of the number of independent samples.

5.2.3 | Combining bias with stability

Due to space constraint, Table 4 presents a selective set of conditions that were especially problematic. First, we walk through the results from one condition to illustrate how to interpret simulation findings. Condition A reflects the situation where mean criterion reliability was .50, the correlation between criterion reliability and true validity was −.50, mean true validity was .100, and true validity *SD* was .300. The estimated mean true validity was .076, thus creating a bias of .076 − .010 = −.024. The *SD* values reflect the variability of estimated mean true validity across 100 iterations with a certain number of independent samples. For example, with 10 independent samples, the *SD* was .108, which suggests that the actual estimate of mean true validity could greatly deviate from the population true value. Note the *SD* values here should not be confused with true validity *SD*. Consistent with what the big *SD* of .108 suggests, 69% of the 100 iterations produced an estimated mean true validity that was outside of .100 ± .05. Interpretations for true validity *SD* results are similar, except there were separate estimated true validity *SD* values under each condition of the number of independent samples. This is because the number of independent samples was also a pertinent factor in influencing its bias. Therefore, results from the four levels are presented separately.

TABLE 4 Simulation results from six conditions (Study 3)

	(A) ρ_{yy} = .50; $\rho_{\rho yy,\ \rho xy}$ = −.50				(B) ρ_{yy} = .90; $\rho_{\rho yy,\ \rho xy}$ = .50				(C) ρ_{yy} = .50; $\rho_{\rho yy,\ \rho xy}$ = −.50			
Statistic	k = 10	k = 30	k = 50	k = 100	k = 10	k = 30	k = 50	k = 100	k = 10	k = 30	k = 50	k = 100
Mean true validity												
$\bar{\rho}_{xy}$	.100				.300				.300			
$\widehat{\bar{\rho}_{xy}}$	.076				.274				.270			
Bias:												
$\widehat{\bar{\rho}_{xy}} - \bar{\rho}_{xy}$	−.024				−.026				−.030			
Stability:												
SD	.108	.048	.051	.034	.085	.055	.040	.029	.095	.058	.039	.028
% outside of $\bar{\rho}_{xy}$ ±.05	69%	38%	35%	22%	52%	41%	34%	22%	62%	45%	33%	26%
True validity SD												
σ_{ρ}	.300				.300				.300			
$\widehat{\sigma_{\rho}}$	.277	.289	.292	.294	.264	.275	.278	.279	.272	.284	.286	.288
Bias:												
$\widehat{\sigma_{\rho}} - \sigma_{\rho}$	−.023	−.011	−.008	−.006	−.036	−.025	−.022	−.021	−.028	−.016	−.014	−.012
Stability:												
SD	.070	.045	.030	.024	.065	.036	.029	.019	.069	.034	.035	.022
% outside of σ_{ρ} ±.05	52%	32%	10%	4%	51%	27%	20%	8%	51%	22%	16%	2%
	(D) ρ_{yy} = .90; $\rho_{\rho yy,\ \rho xy}$ = −.50				(E) ρ_{yy} = .90; $\rho_{\rho yy,\ \rho xy}$ = −.50				(F) ρ_{yy} = .90; $\rho_{\rho yy,\ \rho xy}$ = −.30			
Statistic	k = 10	k = 30	k = 50	k = 100	k = 10	k = 30	k = 50	k = 100	k = 10	k = 30	k = 50	k = 100
Mean true validity												
$\bar{\rho}_{xy}$	.300				.100				.300			
$\widehat{\bar{\rho}_{xy}}$	.313				.120				.304			
Bias:												
$\widehat{\bar{\rho}_{xy}} - \bar{\rho}_{xy}$	.013				.020				.004			
Stability:												
SD	.089	.049	.041	.027	.090	.052	.044	.034	.101	.051	.041	.031
% outside of $\bar{\rho}_{xy}$ ±.05	56%	37%	22%	6%	61%	35%	39%	17%	59%	30%	27%	9%
True validity *SD*												
σ_{ρ}	.300				.300				.300			
$\widehat{\sigma_{\rho}}$	.261	.271	.273	.275	.267	.279	.281	.282	.269	.279	.281	.283
Bias:												
$\widehat{\sigma_{\rho}} - \sigma_{\rho}$	−.039	−.029	−.027	−.025	−.033	−.021	−.019	−.018	−.031	−.021	−.019	−.017
Stability:												
SD	.063	.035	.029	.018	.067	.033	.032	.024	.057	.037	.026	.020
% outside of σ_{ρ} ±.05	54%	29%	22%	11%	52%	21%	18%	8%	50%	27%	13%	4%

Note. ρ_{yy} = mean criterion reliability; $\rho_{\rho yy,\ \rho xy}$ = correlation between criterion reliability and true validity; $\bar{\rho}_{xy}$ = mean true validity; $\widehat{\bar{\rho}_{xy}}$ = estimated mean true validity; $\widehat{\bar{\rho}_{xy}} - \bar{\rho}_{xy}$ = estimation bias of mean true validity; k = number of independent samples; σ_{ρ} = true validity *SD*; $\widehat{\sigma_{\rho}}$ = estimated true validity *SD*; $\widehat{\sigma_{\rho}} - \sigma_{\rho}$ = estimation bias of true validity *SD*; % outside of $\bar{\rho}_{xy}$ ±.05 and % outside of σ_{ρ} ±.05 refer to the percentage of meta-analytic estimates that fell outside of the ±.05 intervals of the true population values of mean true validity and true validity *SD*, respectively.

Next, we sought to evaluate the degree of bias in estimated mean true validity ($\widehat{\bar{\rho}_{xy}} - \bar{\rho}_{xy}$) and its *SD* ($\widehat{\sigma_\rho} - \sigma_\rho$) across different simulated conditions. Clearly, not every condition was problematic with regard to bias. Consistent with OLS results, those that were especially conducive to estimation bias in mean true validity (e.g., Conditions A, B, and C in Table 4) involve a substantial correlation between criterion reliability and true validity (i.e., independence assumption violations). Regarding estimated true validity *SD*, Conditions B, D, E, and F (when $k = 10$) illustrate situations where it may be especially biased. Again, this happened with a strong correlation between criterion reliability and true validity.

Furthermore, as bias only captures the "average" degree of discrepancy across many iterations of meta-analyses,[9] it is important to take into consideration both bias and stability to gain a complete understanding of the applicability of these simulation findings to an individual meta-analysis. We use Condition C to illustrate this issue, where the average bias of mean true validity was the greatest (−.030). Evidently, the lack of stability made this matter worse. When the number of independent samples was only 10, 62 (out of the 100; 62%) meta-analyses yielded an estimate that was either smaller than .25 or greater than .35. Further, increasing the number of independent samples was not able to fully mitigate this issue. Even with 100 independent samples, 26 (out of the 100; 26%) meta-analyses produced an estimate outside the ±.05 interval of the true population value. As evident in this example, although some might think an average bias of −.030 is not that problematic, the actual discrepancy that may manifest in a given meta-analysis can be considerably worse. With 10 independent samples, researchers are more likely to get an estimate outside the ±.05 interval of the true value than to get an estimate within this interval (62%: 38%). Even with 100 independent samples, there is greater than a one-out-of-four chance (26%) that the actual estimate falls outside the ±.05 interval.

On balance, although violating the independence assumptions is not always problematic, it can create a situation where the actual discrepancy in a given meta-analysis is big enough to lead to meaningfully different interpretations. In other words, independence assumption violations are not always a cause for concern, but when they are, they can undermine the accuracy of meta-analytic estimates. Accordingly, we encourage researchers to pay close attention to the detailed results in the Supporting Information, especially the simulated conditions that resemble population parameters in their own research domains. Further, when these results are viewed collectively, the problematic conditions are not a rare occurrence. With 100 independent samples included in a meta-analysis, 27 out of the 75 conditions had a chance of 10% or higher of producing an estimated mean true validity outside of the ±.05 interval around the true population value. With 50 independent samples, 27 out of the 75 conditions had a chance of 10% or higher of yielding an estimated true validity *SD* outside of the ±.05 interval of the true value. Consistent with OLS results supporting the role of independence assumption violations, most of these problematic conditions contained correlated true validity and criterion reliability. Further, it is important to point out that 50 or 100 independent samples are not that common in a published psychometric meta-analysis. For example, the average number of independent samples across the 196 meta-analyses reviewed by Aguinis et al. (2011) was only 18, a level where the lack of stability may be an even greater concern.

6 | GENERAL DISCUSSION

6.1 | Summary of findings

In our research, we sought to first test the tenability of the independence assumptions underlying psychometric meta-analysis and then explore the ramifications of violating such assumptions. Suggesting that the independence between moderator and criterion reliability may not always hold, in Study 1 we found a substantial, nonzero relationship between the moderator of safety climate and safety compliance reliability. In Study 2, our review of recently published psychometric meta-analyses provided further evidence that the violation of independence assumptions is more prevalent than previously assumed. In Study 3, simulation results showed that in certain situations when the independence between criterion reliability and true validity was violated, it could lead to biases in the estimates of both mean true validity and its *SD*. Additionally, if there is a lack of stability, actual estimates from a given meta-analysis may further deviate from the average bias.

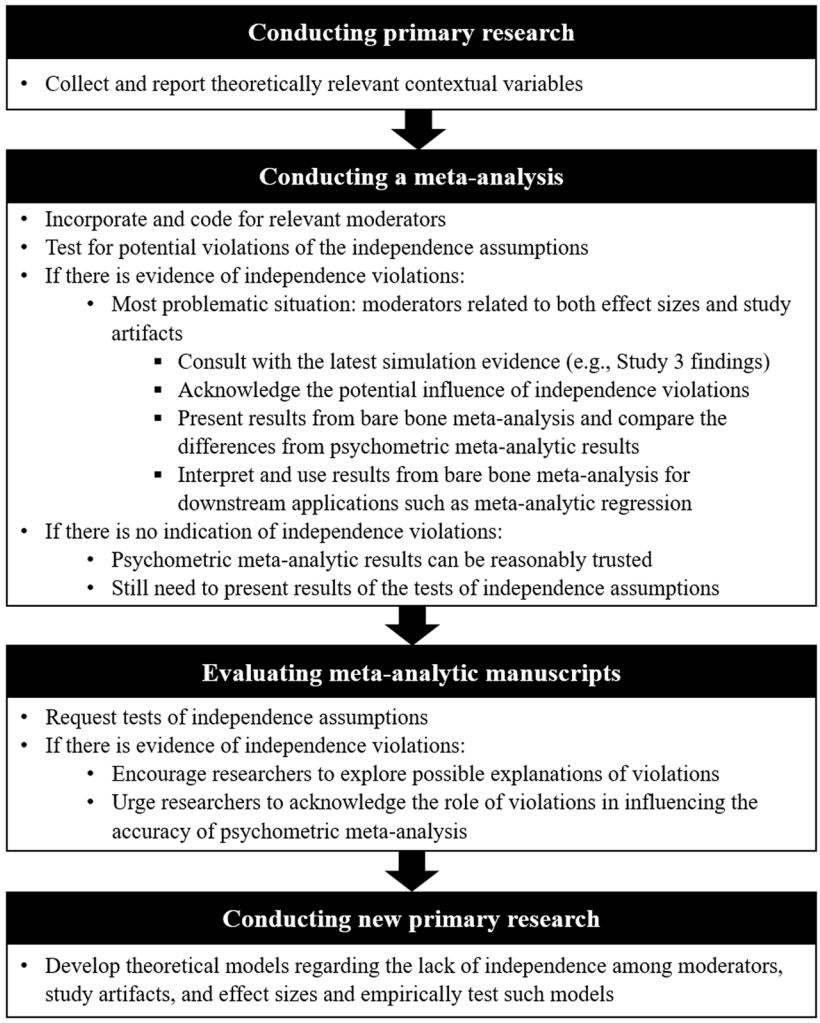

FIGURE 5 Suggestions for researchers, reviewers, and editors

6.2 | Suggestions for psychometric meta-analysis users

Consistent with our findings, we urge researchers to pay close attention to the potential violation of independence assumptions. Incorporating moderators and testing for potential violations will not only promote accurate applications of psychometric meta-analysis, but also drive theoretical advancements. Accordingly, we offer suggestions that are relevant for researchers conducting primary studies and meta-analytic reviews, and editors and reviewers evaluating meta-analyses. These suggestions are summarized in Figure 5.

6.2.1 | Conducting primary research

Researchers conducting primary studies are encouraged to collect data on and report contextual variables whenever it is feasible to do so, especially those that are directly relevant to the research question under investigation. Such detailed reporting will facilitate subsequent meta-analytic efforts examining the roles of moderators and independence assumption violations. The identification of moderators from primary studies plays a much greater role, as it helps to address the limitation of the post hoc nature of moderator detection in psychometric meta-analysis

(Cortina, 2003; James, Demaree, Mulaik, & Ladd, 1992; LeBreton, Schoen, & James, 2017). Specifically, although the logical premise is that there will be a sizable residual variance if moderators exist (if A, then B), psychometric meta-analysis relies on residual variance to infer the existence of moderators (B, therefore A). This is a logical fallacy known as affirming the consequent (Cortina, 2003; James, Demaree, Mulaik, & Ladd, 1992). A problem with relying on this post hoc approach lies in the fact that meaningful moderators that are key to theoretical advancement and refinement "cannot be identified if they are never considered [*a priori*]" (Cortina, 2003, p. 433), or reported in the first place.

6.2.2 | Conducting a meta-analysis

In designing and planning a meta-analytic review, researchers are advised to incorporate and code for relevant moderators by adopting a systematic and consistent operationalization of moderators. Next, researchers should test for potential violations of independence assumptions by inspecting the covariation among study artifacts (i.e., predictor and criterion reliability, range restriction), moderators, and validity coefficients (see Supporting Information for a recommended approach). Whenever feasible, we encourage researchers to conduct the most accurate analysis that takes into account the precision of study artifacts from different samples (e.g., reliability generalization as in Study 1). Nevertheless, product moment correlations can also be informative. To do so, for every meta-analyzed relationship, researchers need to first compile various pieces of information from different independent samples, including the sample size, study artifacts, the effect size, and moderators, and then compute the correlations among these factors. Next, researchers need to effectively summarize these findings in the manuscript by noting any substantial correlations from the correlation analyses and make detailed results regarding testing for independence violations accessible in the supplementary material.

The most problematic situation involves moderators related to both study artifacts and effect sizes. According to James, Demaree, Mulaik, and Ladd (1992), this would suggest study artifacts and effect sizes share a common cause, and therefore are not independent of each other. This constitutes a violation of the independence assumption that could lead to biased results as shown in Study 3. If this situation is identified, researchers should consult with the available simulation evidence and clearly acknowledge how violating the independence assumptions may influence the accuracy of meta-analytic findings. Identifying independence violations also presents an opportunity for future research—a point we return to later. Interpretations based on bare bones meta-analysis (i.e., only correcting for sampling error; no psychometric corrections) are preferred as conservative estimates, especially when results from bare bones meta-analysis are markedly different from those based on problematic psychometric corrections. Further, it is recommended to use results based on bare bones meta-analysis for downstream meta-analytic methods such as meta-analytic structural equation modeling (Yu, Downes, Carter, & O'Boyle, 2016) and meta-analytic regression (Hedges & Olkin, 1985; Steel & Kammeyer-Mueller, 2002), as they both require accurate estimates of mean true validity and its *SD*.

If there is no indication of independence assumption violations, researchers can reasonably trust the findings based on psychometric corrections, provided that plausible moderators have all been identified and tested.[10] Findings regarding the tests of independence assumptions should still be reported for the sake of transparency, which will facilitate future methodological research evaluating the consequences of violating independence assumptions involving other study artifacts.

6.2.3 | Evaluating meta-analytic manuscripts

When evaluating a psychometric meta-analysis, we strongly encourage editors and reviewers to request such tests for violations. There is little harm in performing and reporting these tests. More importantly, when assumptions have been violated, it provides an opportunity for researchers to explore possible explanations for those violations. At a broader level, this will also prove effective in raising researchers' collective awareness of the psychometric meta-analytic assumptions and promoting better executions of this important method. Given that we did not locate any

recent meta-analyses that discussed the role of independence assumptions, we think it is time that editors and reviewers take a proactive part in promoting better understanding and executions of this method.

6.2.4 | Conducting new primary research

Importantly, discoveries of independence assumption violations can guide subsequent exploratory work and ultimately lead to theoretical refinement and advances. In other words, testing for potential independence violations is more than attending to the statistical assumptions of psychometric meta-analysis. It is congruent with the ultimate goal of understanding the effects of contextual variables on workplace phenomena. At the identification of independence violations, we urge researchers to further explore the substantive reasons that moderators influence both effect sizes and study artifacts. Although research has been conducted to understand how moderators can influence effect sizes (e.g., Johns, 2018), we encourage researchers to develop and test plausible linkages between moderators and study artifacts through a systematic examination of contextual influences.

One plausible mechanism is through restrictions in criterion variance (Johns, 2006). For example, James, Demaree, Mulaik, and Ladd (1992) suggest that contextual variables such as organizational climate can discourage the expression of individual differences, thus resulting in restricted criterion variance and lower criterion reliability (see also James, Demaree, & Mulaik, 1992; LeBreton, Burgess, Kaiser, Atchley, & James, 2003). Additionally, other contextual aspects can influence measurement accuracy. For example, high-performance work practices involve using improved performance measurement systems (Takeuchi, Lepak, Wang, & Takeuchi, 2007), suggesting that this contextual variable can act as the substantive cause of higher versus lower criterion reliability across organizations. In addition, job complexity is related to the test–retest reliability of job performance, such that "the greater complexity of the job should actually increase the amount of transient error at any point in time" (Sturman, Cheramie, & Cashen, 2005, p. 275). Moreover, to the extent job complexity influences the correlations among self-, peer-, and supervisor-ratings (e.g., Harris & Schaubroeck, 1988), interrater reliability estimates from 360-degree assessments may be systematically related to the complexity of the job.

These are just a handful of examples illustrating the substantive reasons moderators can be related to study artifacts. However, they do highlight the unique insights that could be gained from this line of inquiry. Specifically, it extends research on contextual influence by revealing its effect on study artifacts, and contributes to a comprehensive understanding of the interplay among contextual variables, effect sizes, and study artifacts. Practically speaking, it informs organizations of how contextual factors at different hierarchical levels of their business environment can influence both the measurement accuracy of personnel testing and the effectiveness of human resource practices. As such, we encourage researchers to explore any potential moderator that may be substantively related to study artifacts and effect sizes in their respective domain. Researchers should first develop plausible hypotheses regarding how moderators may influence study artifacts and effect sizes. Next, they should accurately measure moderators and sample from work environments that vary on the moderators through a large-scale cross-situation study, followed by a formal test of the newly developed hypotheses (James, Demaree, Mulaik, & Ladd, 1992).

6.3 | Limitations and future research directions

There are several limitations associated with the current research worth noting. First, in Study 1, although we controlled for problems associated with different operationalizations of safety performance by including studies utilizing the most commonly used scale, various safety climate scales were used across the primary studies. Therefore, heterogeneity in the operationalization of safety climate may have masked our null finding regarding safety participation (Hunt, 1997). We encourage researchers to conceptualize and operationalize variables consistently in future research along this line. Second, we used coefficient alpha to operationalize criterion reliability in Study 1, as it remains the most commonly reported estimate of reliability used in I/O psychology and management, despite its various limitations (Cortina, 1993; McNeish, 2018; Schmitt, 1996). However, some meta-analyses included in Study 2 used other operationalizations of reliability (e.g., group-mean reliability), which also demonstrated a lack of independence with

other study artifacts and/or moderators. Therefore, the choice of using coefficient alpha is unlikely to fully account for our study findings. Third, most independent samples in Study 1 were based on self reports, thus raising the concern of common method bias (Podsakoff, MacKenzie, & Podsakoff, 2012).[11] However, in Study 2, many of the moderators were not based on self reports (e.g., national culture of the sample; lab versus field studies). Yet, they still demonstrated substantial correlations with reliability. Taken together, common method variance is unlikely to have biased our findings. Relatedly, the lack of independence between moderators and study artifacts could be caused by unmeasured third variables, which merits attention in future research as this would further call into question the independence assumptions in psychometric meta-analysis.[12]

Regarding suggestions for future research, we would like to call for additional methodological research aimed at evaluating the consequences of violating independence assumptions. Specifically, despite the scope of Study 3 (i.e., 7 manipulated factors and 4,800 conditions), we did not consider other artifacts. Future research should incorporate other study artifacts including predictor reliability and range restriction. Moreover, we only considered the situations where the population correlations are normally distributed. Future research can consider nonnormal distributions (Oswald, 1999; Oswald & Johnson, 1998). Additionally, the Type I and II errors of the significance of the population point estimate can also be investigated in the future. In designing simulation studies, we encourage researchers to adopt a similar approach as in Study 3 and anchor their population parameters with the published literature to enhance the applicability of their findings.

7 | CONCLUSION

We conducted this research to evaluate the possibility, prevalence, and consequences of violating independence assumptions in psychometric meta-analysis. Results from a reliability generalization study (Study 1) and a review of recently published meta-analyses (Study 2) suggest violations are not only plausible but also more widespread than previously assumed. Further, simulation results from Study 3 indicate that violating the independence assumptions resulted in biases in certain situations. Further, the lack of stability can produce an actual estimate that substantially deviates from the average bias, thus leading to meaningfully different conclusions. Overall, the role of independence assumptions warrants attention from scholars conducting psychometric meta-analyses, from scholars studying psychometric meta-analysis as a statistical technique, and from consumers of psychometric meta-analysis findings.

ACKNOWLEDGMENTS

The authors thank Jeremy Beus, Michael Christian, and Jennifer Nahrgang for kindly sharing the coding sheets of their published meta-analytic reviews, which helped us identify additional studies for inclusion in Study 1. The authors also want to express gratitude to scholars who responded to their request for unpublished studies to be included in Study 1, and those who conducted additional analyses from their published meta-analyses in Study 2. The authors would also like to thank Michael Burke, Bradley Mecham, Frederick Oswald, Rong Su, Scott Tonidandel, and Qi Zhang for their insights and expertise regarding the design and execution of Monte Carlo simulations in Study 3. Special thanks to Chloe Saucedo for her copy editing assistance. A previous version of this manuscript was presented at the 78th Annual Meeting of the Academy of Management in Chicago, IL.

ENDNOTES

[1] In its original context, validity generalization deals with the question of whether validity coefficients accumulated in previous personnel selection research are generalizable, which typically involves meta-analyzing validity coefficients across different settings (Schmidt & Hunter, 2015). In this paper, we are concerned with the general method of psychometric meta-analysis. Therefore, we use validity generalization to broadly refer to psychometric meta-analysis.

[2] The 53 meta-analyses examined were published in *Academy of Management Journal*, *Journal of Applied Psychology*, *Journal of Management*, and *Personnel Psychology* and excluded meta-analyses on strategic management topics.

[3] We use the general term of moderators to refer to any systematic differences that can differentiate one subpopulation from another, including those that are theoretical and methodological.

[4] Studies included in Study 1: Barbaranelli, Petitta, and Probst (2015); Beus (2012); Beus, Muñoz, and Arthur (2015); Biggs and Banks (2012); Boughaba, Hassane, and Roukia (2014); Braunger, Frank, Korunka, Lueger, and Kubicek (2013); Braunger, Korunka, Kubicek, Frank, and Lueger (2015); Britton (2014); Che (2015); Fleming (2012); Freiwald (2013); Froko, Maxwell, and Kingsley (2015); Gatien (2010); Giffin and Neal (2000); Guros (2015); Hon, Chan, and Yam (2014); Jebb (2015); Mariani, Curcuruto, Matic, Sciacovelli, and Toderi (2017); Martínez-Córcoles and Stephanou (2017); Mashi, Al Subramaniam, and Johari (2017); Mattson, Hellgren, and Göransson (2014); Neal and Griffin (2006); Neal, Griffin, and Hart (2000); Pearce (2012); Petitta, Probst, Barbaranelli, and Ghezzi (2017); Probst (2004); Schwatka (2014); Scott (2016); Shen, Ju, Koh, Rowlinson, and Bridge (2017); Smith and DeJoy (2014); Tang, Leka, Hunt, and MacLennan (2014); Tucker (2010); Xu et al. (2014); Yuan (2014); and Yuan, Li, and Tetrick (2015).

[5] A close inspection of these two samples showed that they had either extremely low criterion reliability or mean safety climate (Braunger, Frank, Korunka, Lueger, & Kubicek, 2013; Guros, 2015).

[6] Meta-analyses included in Study 2: Banks, Woznyj, Kepes, Batchelor, and McDaniel (2018); Beus, Dhanani, and McCord (2015); Casper, Vaziri, Wayne, DeHauw, and Greenhaus (2018); Chamberlin, Newton, and Lepine (2017); Choi, Oh, and Colbert (2015); Courtright, Thurgood, Stewart, and Pierotti (2015); De Jong, Dirks, and Gillespie (2016); Greco, O'Boyle, and Walter (2015); Joseph, Jin, Newman, and O'Boyle (2015); Knight and Eisenkraft (2015); Litwiller, Snyder, Taylor, and Steele (2017); Mackey, Frieder, Brees, and Martinko (2017); Martin, Guillaume, Thomas, Lee, and Epitropaki (2016); McCord, Joseph, Dhanani, and Beus (2018); Morris, Daisley, Wheeler, and Boyer (2015); Nohe, Meier, Sonntag, and Michel (2015); and Van den Broeck, Ferris, Chang, and Rosen (2016).

[7] In this process, it is possible that criterion reliability and/or true validity may assume out of bound values. Accordingly, we first drew a sufficiently large number of sets of values and kept the first *k* set of values that were within the normal bound of reliability (between 0 and 1) and true validity (between −1 and 1).

[8] This was only available in the development version of "metafor" when we were conducting this simulation study (see Viechtbauer, 2019).

[9] In the case of mean true validity, each of the 75 bias values reflects the average bias based on 64,000 iterations of psychometric meta-analysis, as we collapsed the estimates across nonsignificant manipulated factors. Regarding true validity *SD*, each bias value was the average from 16,000 iterations.

[10] Even in this case, we would like to note that estimates from a small number of independent studies are likely unstable (see Table 4 and the Supporting Information).

[11] In Study 1, Braunger, Korunka, Kubicek, Frank, and Lueger (2015) and Xu et al. (2014) used non-self-reported safety performance ratings.

[12] We thank an anonymous reviewer for alerting us to this point.

ORCID

Zhenyu Yuan https://orcid.org/0000-0002-7971-887X
Frederick P. Morgeson https://orcid.org/0000-0001-6858-2594
James M. LeBreton https://orcid.org/0000-0001-6748-380X

REFERENCES

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

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*, 270–301. https://doi.org/10.1177/1094428112470848

Andriessen, J. H. T. H. (1978). Safe behavior and safety motivation. *Journal of Occupational Accidents*, *1*, 363–376. https://doi.org/10.1016/0376-6349(78)90006-8

Banks, G. C., Woznyj, H. M., Kepes, S., Batchelor, J. H., & McDaniel, M. A. (2018). A meta-analytic review of tipping compensation practices: An agency theory perspective. *Personnel Psychology*, *71*, 457–478. https://doi.org/10.1111/peps.12261

Barbaranelli, C., Petitta, L., & Probst, T. M. (2015). Does safety climate predict safety performance in Italy and the USA? Cross-cultural validation of a theoretical model of safety climate. *Accident Analysis & Prevention*, *77*, 35–44. https://doi.org/10.1016/j.aap.2015.01.012

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Beus, J. M. (2012). The psychological need for safety at work: A cybernetic perspective (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.

Beus, J. M., Dhanani, L. Y., & McCord, M. A. (2015). A meta-analysis of personality and workplace safety: Addressing unanswered questions. *Journal of Applied Psychology*, *100*, 481–498. https://doi.org/10.1037/a0037916

Beus, J. M., Muñoz, G. J., & Arthur, W. (2015). Personality as a multilevel predictor of climate: An examination in the domain of workplace safety. *Group & Organization Management*, *40*, 625–656. https://doi.org/10.1177/1059601115576597

Biggs, S. E., & Banks, T. D. (2012, September 20–21). *A comparison of safety climate and safety outcomes between construction and resource functions in a large case study organisation*. Paper presented at Occupational Safety in Transport Conference, Gold Coast, Queensland.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. https://doi.org/10.1037/a0038047

Boughaba, A., Hassane, C., & Roukia, O. (2014). Safety culture assessment in petrochemical industry: A comparative study of two Algerian plants. *Safety and Health at Work*, *5*, 60–65. https://doi.org/10.1016/j.shaw.2014.03.005

Braunger, P., Frank, H., Korunka, C., Lueger, M., & Kubicek, B. (2013). Validating a safety climate model in metal processing industries: A replication study. *International Journal of Occupational Safety and Ergonomics*, *19*, 143–155. https://doi.org/10.1080/10803548.2013.11076973

Braunger, P., Korunka, C., Kubicek, B., Frank, H., & Lueger, M. (2015). The perspective of safety engineers on safety climate. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *25*, 198–210. https://doi.org/10.1002/hfm.20538

Britton, A. R. (2014). Safety-specific person-environment fit: Relation with safety behaviors, job attitudes, and strain (Unpublished doctoral dissertation). Bowling Green State University, Bowling Green, OH.

Burke, M. J., Rupinski, M. T., Dunlap, W. P., & Davison, H. K. (1996). Do situational variables act as substantive causes of relationships between individual difference variables? Two large-scale tests of "common cause" models. *Personnel Psychology*, *49*, 573–598. https://doi.org/10.1111/j.1744-6570.1996.tb01585.x

Casper, W. J., Vaziri, H., Wayne, J. H., DeHauw, S., & Greenhaus, J. (2018). The jingle-jangle of work–nonwork balance: A comprehensive and meta-analytic review of its meaning and measurement. *Journal of Applied Psychology*, *103*, 182–214. https://doi.org/10.1037/apl0000259

Chamberlin, M., Newton, D. W., & Lepine, J. A. (2017). A meta-analysis of voice and its promotive and prohibitive forms: Identification of key associations, distinctions, and future research directions. *Personnel Psychology*, *70*, 11–71. https://doi.org/10.1111/peps.12185

Che, X. (2015). Effects of occupational stressors on nurses' safety performance and well-being: A within-individual study (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.

Choi, D., Oh, I. S., & Colbert, A. E. (2015). Understanding organizational commitment: A meta-analytic examination of the roles of the five-factor model of personality and culture. *Journal of Applied Psychology*, *100*, 1542–1567. https://doi.org/10.1037/apl0000014

Christian, M. S., Bradley, J. C., Wallace, J. C., & Burke, M. J. (2009). Workplace safety: A meta-analysis of the roles of person and situation factors. *Journal of Applied Psychology*, *94*, 1103–1127. https://doi.org/10.1037/a0016172

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods*, *6*, 415–439. https://doi.org/10.1177/1094428103257358

Courtright, S. H., Thurgood, G. R., Stewart, G. L., & Pierotti, A. J. (2015). Structural interdependence in teams: An integrative framework and meta-analysis. *Journal of Applied Psychology*, *100*, 1825–1846. https://doi.org/10.1037/apl0000027

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology*, *101*, 1134–1150. https://doi.org/10.1037/apl0000110

De Wit, F. R., Greer, L. L., & Jehn, K. A. (2012). The paradox of intragroup conflict: A meta-analysis. *Journal of Applied Psychology*, *97*, 360–390. https://doi.org/10.1037/a0024844

Fleming, M. (2012). Assessing employee safety motivation (Working Paper No. RS2010-DG08). Retrieved from http://www.worksafebc.com/contact_us/research/funding_decisions/assets/pdf/2012/RS2010-DG08.pdf

Freiwald, D. R. (2013). The effects of ethical leadership and organizational safety culture on safety outcomes (Unpublished doctoral dissertation). Embry-Riddle Aeronautical University, Daytona Beach, FL.

Froko, I. U., Maxwell, A., & Kingsley, N. (2015). The impact of safety climate on safety performance in a gold mining company in Ghana. *International Journal of Management Excellence*, *5*, 556–566.

Gatien, B. (2010). An investigation into the relationship between perceptions of safety climate and organizational justice (Unpublished master's thesis). Saint Mary's University, Halifax, Canada.

Gillespie, M. A., Oswald, F. L., & Converse, P. D. (2002). On using meta-analysis to make judgments about validity generalization. In S. Morris (Chair), Rethinking artifact corrections in meta-analysis: Innovations and extensions. Symposium presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, CN.

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies*, *55*, 583–618. https://doi.org/10.1111/joms.12328

Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2015). Absence of malice: A meta-analysis of nonresponse bias in counterproductive work behavior research. *Journal of Applied Psychology*, *100*, 75–97. https://doi.org/10.1037/a0037495

Griffin, M. A., & Neal, A. (2000). Perceptions of safety at work: A framework for linking safety climate to safety performance, knowledge, and motivation. *Journal of Occupational Health Psychology*, *5*, 347–358. https://doi.org/10.1037/1076-8998.5.3.347

Guros, F. (2015). Thinking about work at home: Implications for safety at work (Unpublished doctoral dissertation). Portland State University, Portland, OR.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*, 377–389. https://doi.org/10.1037/0021-9010.87.2.377

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41*, 43–62. https://doi.org/10.1111/j.1744-6570.1988.tb00631.x

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hofmann, D. A., Morgeson, F. P., & Gerras, S. J. (2003). Climate as a moderator of the relationship between leader-member exchange and content specific citizenship: Safety climate as an exemplar. *Journal of Applied Psychology*, *88*, 170–178. https://doi.org/10.1037/0021-9010.88.1.170

Hon, C. K., Chan, A. P., & Yam, M. C. (2014). Relationships between safety climate and safety performance of building repair, maintenance, minor alteration, and addition (RMAA) works. *Safety Science*, *65*, 10–19. https://doi.org/10.1016/j.ssci.2013.12.012

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York, NY: Russell Sage Found.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*, 340–362. https://doi.org/10.1016/0001-8791(86)90013-8

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98. https://doi.org/10.1037/0033-2909.96.1.72

James, L. R., Demaree, R. G., & Mulaik, S. A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology*, *71*, 440–450. https://doi.org/10.1037/0021-9010.71.3.440

James, L. R., Demaree, R. G., & Mulaik, S. A. (1992). A critique of validity generalization. In B. R. Gifford & L. C. Wing (Eds.), *Report of the National Commission on Testing and Public Policy* (pp. 13–76). Boston, MA: Kluwer-Nijhof.

James, L. R., Demaree, R. G., Mulaik, S. A., & Ladd, R. T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology*, *77*, 3–14. https://doi.org/10.1037/0021-9010.77.1.3

James, L. R., Demaree, R. G., Mulaik, S. A., & Mumford, M. D. (1988). Validity generalization: Rejoinder to Schmidt, Hunter, and Raju. *Journal of Applied Psychology*, *73*, 673–678. https://doi.org/10.1037/0021-9010.73.4.673

Jebb, S. E. (2015). Reducing workplace safety incidents: Bridging the gap between safety culture theory and practice (Unpublished doctoral dissertation). Queensland University of Technology, Brisbane, Australia.

Jiang, L., Lavaysse, L. M., & Probst, T. M. (2019). Safety climate and safety outcomes: A meta-analytic comparison of universal vs. industry-specific safety climate predictive validity. *Work & Stress*, *33*, 41–57. https://doi.org/10.1080/02678373.2018.1457737

Jiang, L., Yu, G., Li, Y., & Li, F. (2010). Perceived colleagues' safety knowledge/behavior and safety performance: Safety climate as a moderator in a multilevel study. *Accident Analysis & Prevention*, *42*, 1468–1476. https://doi.org/10.1016/j.aap.2009.08.017

Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review*, *31*, 386–408. https://doi.org/10.5465/amr.2006.20208687

Johns, G. (2018). Advances in the treatment of context in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*, 21–46. https://doi.org/10.1146/annurev-orgpsych-032117-104406

Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology*, *100*, 298–342. https://doi.org/10.1037/a0037681

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—Article, author, or journal? *Academy of Management Journal*, *50*, 491–506. https://doi.org/10.5465/amj.2007.25525577

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction-job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*, 376–407. https://doi.org/10.1037/0033-2909.127.3.376

Kemery, E. R., Mossholder, K. W., & Roth, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology*, *72*, 30–37. https://doi.org/10.1037/0021-9010.72.1.30

Knight, A. P., & Eisenkraft, N. (2015). Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance. *Journal of Applied Psychology*, *100*, 1214–1227. https://doi.org/10.1037/apl0000006

Köhler, T., Cortina, J. M., Kurtessis, J. N., & Gölz, M. (2015). Are we correcting correctly? Interdependence of reliabilities in meta-analysis. *Organizational Research Methods*, *18*, 355–428. https://doi.org/10.1177/1094428114563617

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology*, *79*, 978–986. https://doi.org/10.1037/0021-9010.79.6.978

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*, 80–128. https://doi.org/10.1177/1094428102239427

LeBreton, J. M., Schoen, J. L., & James, L. R. (2017). Situational specificity, validity generalization, and the future of psychometric meta-analysis. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 93–114). New York, NY: Routledge.

Litwiller, B., Snyder, L. A., Taylor, W. D., & Steele, L. M. (2017). The relationship between sleep and work: A meta-analysis. *Journal of Applied Psychology*, *102*, 682–699. https://doi.org/10.1037/apl0000169

Mackey, J. D., Frieder, R. E., Brees, J. R., & Martinko, M. J. (2017). Abusive supervision: A meta-analysis and empirical review. *Journal of Management*, *43*, 1940–1965. https://doi.org/10.1177/0149206315573997

Marchand, A., Simard, M., Carpentier-Roy, M. C., & Ouellet, F. (1998). From a unidimensional to a bidimensional concept and measurement of workers' safety behavior. *Scandinavian Journal of Work, Environment & Health*, *24*, 293–299. https://doi.org/10.5271/sjweh.323

Mariani, M. G., Curcuruto, M., Matic, M., Sciacovelli, P., & Toderi, S. (2017). Can leader–member exchange contribute to safety performance in an Italian warehouse? *Frontiers in Psychology*, *8*, 1–9. https://doi.org/10.3389/fpsyg.2017.00729

Martin, R., Guillaume, Y., Thomas, G., Lee, A., & Epitropaki, O. (2016). Leader–member exchange (LMX) and performance: A meta-analytic review. *Personnel Psychology*, *69*, 67–121. https://doi.org/10.1111/peps.12100

Martínez-Córcoles, M., & Stephanou, K. (2017). Linking active transactional leadership and safety performance in military operations. *Safety Science*, *96*, 93–101. https://doi.org/10.1016/j.ssci.2017.03.013

Mashi, M. S., Al Subramaniam, C., & Johari, J. B., (2017). The effect of management commitment, safety rules and procedure and safety promotion policies on nurses safety performance: The moderating role of consideration of future safety consequences. *International Business Management*, *100*, 478–489.

Mattson, M., Hellgren, J., & Göransson, S. (2014, June 11–13). Effects of different leader communication strategies on safety behaviors and safety outcomes. Paper presented at the 7th Nordic Working Life Conference, Göteborg, Sweden.

McCord, M. A., Joseph, D. L., Dhanani, L. Y., & Beus, J. M. (2018). A meta-analysis of sex and race differences in perceived workplace mistreatment. *Journal of Applied Psychology*, *103*, 137–163. https://doi.org/10.1037/apl0000250

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433. https://doi.org/10.1037/met0000144

Mendoza, J. L., & Reinhardt, R. N. (1991). Validity generalization procedures using sample-based estimates: A comparison of six procedures. *Psychological Bulletin*, *110*, 596–610. https://doi.org/10.1037/0033-2909.110.3.596

Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology*, *100*, 5–20. https://doi.org/10.1037/a0036938

Nahrgang, J. D., Morgeson, F. P., & Hofmann, D. A. (2011). Safety at work: A meta-analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. *Journal of Applied Psychology*, *96*, 71–94. https://doi.org/10.1037/a0021484

Neal, A., & Griffin, M. A. (2006). A study of the lagged relationships among safety climate, safety motivation, safety behavior, and accidents at the individual and group levels. *Journal of Applied Psychology*, *91*, 946–953. https://doi.org/10.1037/0021-9010.91.4.946

Neal, A., Griffin, M. A., & Hart, P. M. (2000). The impact of organizational climate on safety climate and individual behavior. *Safety Science*, *34*, 99–109. https://doi.org/10.1016/S0925-7535(00)00008-4

Nohe, C., Meier, L. L., Sonntag, K., & Michel, A. (2015). The chicken or the egg? A meta-analysis of panel studies of the relationship between work–family conflict and strain. *Journal of Applied Psychology*, *100*, 522–536. https://doi.org/10.1037/a0038012

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Oswald, F. L. (1999). *On deriving validity generalization and situational specificity from meta-analysis: A conceptual review and some empirical findings* (Unpublished doctoral dissertation). University of Minnesota, Twin Cities, MN.

Oswald, F. L., & Johnson, J. W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology*, *83*, 164–178. https://doi.org/10.1037/0021-9010.83.2.164

Pearce, M. N. (2012). Safety climate, safety behaviours and control: An application of the Job Demand-Control model to occupational safety (Unpublished master's theses). University of Canterbury, Christchurch, New Zealand.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, *65*, 373–406. https://doi.org/10.1037/0021-9010.65.4.373

Petitta, L., Probst, T. M., Barbaranelli, C., & Ghezzi, V. (2017). Disentangling the roles of safety climate and safety culture: Multilevel effects on the relationship between supervisor enforcement and safety compliance. *Accident Analysis & Prevention*, *99*, 77–89. https://doi.org/10.1016/j.aap.2016.11.012

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Podsakoff, P. M., Podsakoff, N. P., Mishra, P., & Escue, C. (2018). Can early-career scholars conduct impactful research? Playing "small ball" versus "swinging for the fences". *Academy of Management Learning & Education*, *17*, 496–531. https://doi.org/10.5465/amle.2017.0198

Probst, T. M. (2004). Safety and insecurity: Exploring the moderating effect of organizational safety climate. *Journal of Occupational Health Psychology*, *9*, 3–10. https://doi.org/10.1037/1076-8998.9.1.3

Raju, N. S., Anselmi, T. V., Goodman, J. S., & Thomas, A. (1998). The effect of correlated artifacts and true validity on the accuracy of parameter estimation in validity generalization. *Personnel Psychology*, *51*, 453–465. https://doi.org/10.1111/j.1744-6570.1998.tb00733.x

Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology*, *68*, 382–395. https://doi.org/10.1037/0021-9010.68.3.382

Raju, N. S., Pappas, S., & Williams, C. P. (1989). An empirical Monte Carlo test of the accuracy of the correlation, covariance, and regression slope models for assessing validity generalization. *Journal of Applied Psychology*, *74*, 901–911. https://doi.org/10.1037/0021-9010.74.6.901

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322. https://doi.org/10.1037/1082-989X.11.3.306

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540. https://doi.org/10.1037/0021-9010.62.5.529

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology*, *33*, 705–724. https://doi.org/10.1111/j.1744-6570.1980.tb02364.x

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Schneider, B., González-Romá, V., Ostroff, C., & West, M. A. (2017). Organizational climate and culture: Reflections on the history of the constructs in the *Journal of Applied Psychology*. *Journal of Applied Psychology*, *102*, 468–482. https://doi.org/10.1037/apl0000090

Schwatka, N. V. (2014). Managing through measurement: Occupational health and safety in the construction industry (Unpublished doctoral dissertation). Colorado State University, Fort Collins, CO.

Scott, N. (2016). Enjoyment, values, pressure, or something else: What influences employees' safety behaviors? (Unpublished doctoral dissertation). Saint Mary's University, Halifax, Canada.

Shen, Y., Ju, C., Koh, T. Y., Rowlinson, S., & Bridge, A. J. (2017). The impact of transformational leadership on safety climate and individual safety behavior on construction sites. *International Journal of Environmental Research and Public Health*, *14*, 45. https://doi.org/10.3390/ijerph14010045

Smith, T. D., & DeJoy, D. M. (2014). Safety climate, safety behaviors and line-of-duty injuries in the fire service. *International Journal of Emergency Services*, *3*, 49–64. https://doi.org/10.1108/IJES-04-2013-0010

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*, 96–111. https://doi.org/10.1037/0021-9010.87.1.96

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology*, *90*, 269–283. https://doi.org/10.1037/0021-9010.90.2.269

Takeuchi, R., Lepak, D. P., Wang, H., & Takeuchi, K. (2007). An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *Journal of Applied Psychology*, *92*, 1069–1083. https://doi.org/10.1037/0021-9010.92.4.1069

Tang, J. J., Leka, S., Hunt, N., & MacLennan, S. (2014). An exploration of workplace social capital as an antecedent of occupational safety and health climate and outcomes in the Chinese education sector. *International Archives of Occupational and Environmental Health*, *87*, 515–526. https://doi.org/10.1007/s00420-013-0890-9

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*, 2693–2708. https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V

Tucker, J. J. (2010). The moderating effect of safety climate on the relationship between job insecurity and employee safety outcomes (Unpublished master's thesis). The University of Wisconsin–Oshkosh, Oshkosh, WI.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. https://doi.org/10.1177/0013164498058001002

Van den Broeck, A., Ferris, D. L., Chang, C. H., & Rosen, C. C. (2016). A review of self-determination theory's basic psychological needs at work. *Journal of Management*, *42*, 1195–1229. https://doi.org/10.1177/0149206316632058

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2019). *Hunter and Schmidt method [Web log post]*. Retrieved from http://www.metafor-project.org/doku.php/tips:hUnter_schmidt_method

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, *75*, 315–321. https://doi.org/10.1037/0021-9010.75.3.315

Xu, Y., Li, Y., Wang, G., Yuan, X., Ding, W., & Shen, Z. (2014). Attentional bias toward safety predicts safety behaviors. *Accident Analysis & Prevention*, *71*, 144–153. https://doi.org/10.1016/j.aap.2014.05.013

Yu, J. J., Downes, P. E., Carter, K. M., & O'Boyle, E. H. (2016). The problem of effect size heterogeneity in meta-analytic structural equation modeling. *Journal of Applied Psychology*, *101*, 1457–1473. https://doi.org/10.1037/apl0000141

Yuan, Z. (2014). A preliminary development and validation of a measure of safety performance (Unpublished master's thesis). Purdue University, Lafayette, IN.

Yuan, Z., Li, Y., & Tetrick, L. E. (2015). Job hindrances, job resources, and safety performance: The mediating role of job engagement. *Applied Ergonomics*, *51*, 163–171. https://doi.org/10.1016/j.apergo.2015.04.021

Zohar, D. (1980). Safety climate in industrial organizations: Theoretical and applied implications. *Journal of Applied Psychology*, *65*, 96–102. https://doi.org/10.1037/0021-9010.65.1.96

Zohar, D. (2000). A group-level model of safety climate: Testing the effect of group climate on microaccidents in manufacturing jobs. *Journal of Applied Psychology*, *85*, 587–596. https://doi.org/10.1037/0021-9010.85.4.587

Zohar, D. (2002). Modifying supervisory practices to improve subunit safety: A leadership-based intervention model. *Journal of Applied Psychology*, *87*, 156–163. https://doi.org/10.1037/0021-9010.87.1.156

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Yuan Z, Morgeson FP, LeBreton JM. Maybe not so independent after all: The possibility, prevalence, and consequences of violating the independence assumptions in psychometric meta-analysis. *Personnel Psychology*. 2020;73:491–516. https://doi.org/10.1111/peps.pep12372