Received: 1 November 2018 | Revised: 3 August 2019 | Accepted: 18 December 2019

DOI: 10.1111/pere.12309

ARTICLE

Personal RELATIONSHIPS WILEY

Solutions to the problems of incremental validity testing in relationship science

Y. Andre Wang | **Paul W. Eastwick**

Department of Psychology, University of California, Davis, California

Correspondence
Y. Andre Wang, Department of Psychology, University of California, One Shields Avenue, Davis, CA 95616.
Email: ylawang@ucdavis.edu

Abstract

Incremental validity testing (i.e., testing whether a focal predictor is associated with an outcome above and beyond a covariate) is common (e.g., 57% of *Personal Relationships* articles in 2017), yet it is fraught with conceptual and statistical problems. First, researchers often use it to overemphasize the novelty or counterintuitiveness of findings, which hinders cumulative understanding. Second, incremental validity testing requires that the focal predictor and the covariate represent separate constructs; researchers risk committing the "jangle fallacy" without such evidence. Third, the most common approach to incremental validity testing (i.e., standard multiple regression, 88% of articles) inflates Type I error and can produce invalid conclusions. This article also discusses the relevance of these issues to dyadic/longitudinal designs and offers concrete solutions.

1 | INTRODUCTION

Relationship researchers often want to know whether a predictor accounts for variance in a dependent variable above and beyond a third, partially overlapping construct. Does satisfaction predict commitment above and beyond desirability of alternatives (Le & Agnew, 2003)? Does relationship status predict partner-specific attachment anxiety above and beyond dispositional attachment anxiety (Eastwick & Finkel, 2008)? Does trust predict fluctuations in relationship quality above and beyond overall relationship quality (Campbell, Simpson, Boldry, & Rubin, 2010)? When examining questions such as these, researchers often attempt to demonstrate *incremental validity*: Is the effect of a focal predictor on an outcome variable significant over and above the third variable (i.e., a covariate)? Tests of incremental validity

are common in relationship science: In the 56 empirical articles published in *Personal Relationships* in 2017, 32 (57%) of them included at least one test of incremental validity (see Table S1 of Supplemental Materials for details). It is safe to say that incremental validity tests (and interpretations thereof) have substantively informed the field of close relationships.

Incremental validity testing arose from the practical and reasonable suggestion that novel predictive measures (e.g., new clinical or personnel tests) should produce an increase in "predictive efficiency over the information otherwise easily and cheaply available" (Sechrest, 1963, p. 154; for a history, see Hunsley & Meyer, 2003). Translated to the context of contemporary relationship research, a demonstration of incremental validity would require that the association between a focal predictor and an outcome goes beyond a well-established (i.e., "easily and cheaply available") association between a covariate and the outcome. Thus, incremental validity is a useful way of demonstrating the empirical novelty of a conceptual contribution that simultaneously acknowledges a precedent established in earlier work. Given that a version of this test can be calculated with widely available multiple regression analytic programs, relationship scholars habitually conduct incremental validity tests in most of their lines of research.

When implemented carefully and thoughtfully, incremental validity testing can contribute to scholars' understanding of the relations between a focal predictor, a covariate, and an outcome variable. As we discuss in the section "How to Test Incremental Validity Convincingly: An Example" below, there are certainly cases where scholars will want to make incremental validity a focal and central component of their scientific claims. However, the test has become sufficiently routinized that incremental validity tests are often disconnected from a strong theoretical rationale. Furthermore, recent psychometric research has highlighted that many incremental validity claims—especially when they use standard multiple regression analyses—may not be statistically valid (Culpepper & Aguinis, 2011; Shear & Zumbo, 2013; Westfall & Yarkoni, 2016).

This article is about *when* scholars should make incremental validity claims and *how* they can do so appropriately. It is structured around three problems with the existing practice of incremental validity testing in close relationship research (Table 1). First, although incremental validity testing is commonplace, when researchers wish to demonstrate that an effect is novel and/or surprising, these tests actually hinder cumulative understanding under many circumstances (Giner-Sorolla, 2012). Second, although incremental validity testing rests on the assumption that the predictor and covariate are separate constructs, it is easy for well-meaning researchers to incorrectly presume that constructs with two different names are in fact two different things (i.e., the *jangle fallacy*; Kelley, 1927). Finally, although there are tests of incremental validity that account for measurement error, the predominant regression approach of testing incremental validity does not, which may greatly inflate the likelihood of committing Type I errors (Culpepper & Aguinis, 2011; Shear & Zumbo, 2013; Westfall & Yarkoni, 2016). These problems are not unique to incremental validity testing, but all three must be addressed if relationship researchers are to use this procedure to generate robust findings and make reliable theoretical progress. To this end, after we review each problem, we offer solutions that should aid researchers in strengthening the way in which they implement the practice of incremental validity testing.

TABLE 1 Summary of potential problems with common approaches to testing incremental validity in relationship research and proposed solutions

Potential problems	Description	Proposed solutions
You might overemphasize novelty and counterintuitiveness.	You might be using incremental validity testing to address an inappropriate application of the "isn't it just...?" critique.	Interrogate any suggestion that the finding lacks value without evidence of novelty and appreciate that counterintuitive explanations are not inherently more valuable than intuitive ones.
You might not have separate constructs.	Your focal predictor (X) and covariate (C) might not be empirically distinct.	Formally test if you have evidence of separate constructs by using confirmatory factor analysis.
Your result might be a type I error.	You will likely increase your Type I error rate—perhaps dramatically—using the standard multiple regression approach.	Use structural equation modeling (SEM) with latent variables to account for measurement error and control your Type I error rate.

2 | INCREMENTAL VALIDITY PROBLEM #1: YOU MIGHT OVEREMPHASIZE NOVELTY AND COUNTERINTUITIVENESS

Sometimes, a theory or model in relationship science clearly delineates the need for an incremental validity test. The Investment Model of Commitment (Rusbult, 1980) offers a case in point. According to this model, three factors (i.e., satisfaction, desirability of alternatives, and investments) independently predict commitment, and so, incremental validity tests (i.e., tests of the predictive association of investments with commitment above and beyond satisfaction and alternatives) are appropriate (Le & Agnew, 2003; Rusbult, 1983). However, in some cases, researchers might feel the need to conduct an incremental validity test not because the test clearly follows from a theory or model but because it heads off the infamous "isn't it just...?" critique.

2.1 | "Isn't it just...?"

Nearly all relationship scholars have encountered the "isn't it just...?" critique. Here is how it often manifests: When a researcher discovers a significant association of a predictor X with an outcome Y, a critic responds with "isn't it just (associated covariate) C?" Researchers defending their X–Y association would typically go on to conduct tests of incremental validity to demonstrate that the X–Y association is not due to one or more covariate Cs. This practice is commonplace; the second author of the current article has done it many times (e.g., testing whether attachment bond strength produces effects above and beyond satisfaction or commitment, Eastwick & Finkel, 2012).

This critique is easy to levy—perhaps especially so in relationship science because of the nature of the variables that close relationship scholars tend to assess. Specifically, close

relationship researchers frequently assess evaluative, affect-laden constructs in their studies, such as relationship satisfaction, passion, love, and commitment. These evaluative constructs are not identical to each other, but they are all influenced by relationship quality and related constructs that reflect the extent to which one feels positively versus negatively about one's partner and relationship (Fletcher, Simpson, & Thomas, 2000). In other words, most evaluative constructs will be associated with each other—and become viable covariates (i.e., Cs)—because they are all themselves influenced by relationship quality. A related, well-known phenomenon in relationship science is called *sentiment override* (Weiss, 1980): Participants tend to respond to questions about specific situations (e.g., "my spouse will not listen fully to what I am saying") by drawing on their predominant positive versus negative sentiment about the partner (i.e., relationship quality; Fincham, Garnier, Gano-Phillips, & Osborne, 1995). Therefore, even precise, specific questionnaire items and behavioral assessments may reflect relationship quality.

When scholars in relationship science use any of the correlated evaluative constructs that potentially tap different elements of relationship quality, critics can implement the "isn't it just...?" critique simply by nominating relationship quality or a related construct as the covariate C. That is, if someone shows that passion (X) predicts the effective provision of support behaviors (Y), the alternative explanatory construct (C) could be relationship quality or any other evaluative construct that is influenced by relationship quality (satisfaction, commitment, etc.). However, before embarking on an incremental validity test to rebut this critique, it is worth pausing to ask what "isn't it just...?" really means in this context. We argue that, at a deep level, "isn't it just...?" can have two meanings, both of which can have insidious, unintended consequences if unchallenged.

2.2 | Meaning #1: If it's just relationship quality, your finding is not novel

One common implicit meaning of "isn't it just...?" is that the finding has already been published. That is, if relationship quality (covariate C) is truly the source of the X–Y association, then the finding is not new and is not worth publishing a second time (i.e., "we knew this already"; Madden, Easley, & Dunn, 1995; Neuliep & Crandall, 1990, 1993).

Novelty—that is, the discovery of new empirical phenomena—is an important scientific goal (Finkel, Eastwick, & Reis, 2017), but an overemphasis on novelty can detract from other, equally essential goals: For example, the "isn't it just...?" critique potentially hinders the goal of building a replicable relationship science because it presumes that published findings are both true and sufficient for understanding a given effect. Indeed, relationship science is unlikely to be immune to replicability challenges (Campbell, Loving, & LeBel, 2014; Cheung et al., 2016; LeBel, Campbell, & Loving, 2017). Therefore, the fact that a particular association between relationship quality (i.e., C) and X or relationship quality and Y has been published is not a sufficient reason to devalue additional tests of it. Furthermore, the increased emphasis on replicability has underscored the importance of cumulative evidence in scientific truth-seeking (Ledgerwood, 2014). Contrary to the implications of "isn't it just...?," a finding with published precedent advances scientific understanding in multiple ways: Not only does it provide more evidence for that effect (i.e., it "replicates" successfully), as well as evidence for its generalizability (i.e., it emerges in a new kind of sample), it also opens up possibilities for future efforts to more precisely and accurately estimate the size of the effect (e.g., via meta-analysis). In this

way, a reduction in the traditional emphasis on novelty can pave the way for scholars to emphasize other essential scientific features such as generalizability and cumulativeness (Finkel et al., 2017). In short, there are many ways in which scholars can emphasize how their X–Y finding bolsters a previously documented association of relationship quality with X or Y, thus highlighting the benefits of their work for replicability, generalizability, and cumulativeness in lieu of novelty.

2.3 | Meaning #2: If it's just relationship quality, your finding is intuitive

A second implicit meaning of "isn't it just...?" can apply even in cases where novelty is not in question (i.e., the association of relationship quality with X or Y have not been published previously). This meaning invokes the idea that relationship processes are intuitive, and perhaps even boring, when they are a function of relationship quality. For example, it is intuitive that relationship quality predicts the effective provision of support behaviors, but it is less intuitive if something far more specific (e.g., "having good sex the night before") predicts the effective provision of support behaviors. If a preference for novelty underlies the first meaning of "isn't it just...?," a preference for counterintuitiveness underlies the second meaning. Importantly, in contrast to novelty, counterintuitiveness is not a goal of a healthy scientific discipline (Kruglanski, Chernikova, & Jasko, 2016; Ledgerwood & Sherman, 2012); there is no scientific imperative to accumulate surprising, "man bites dog" findings.

Results may seem more counterintuitive in relationship research to the extent that the implicated constructs are specific rather than global. This bias is related to the perennial contrast between "lumping" and "splitting"; relationship quality is a lump that obviously predicts many outcomes, but by splitting that lump into finer and finer parts, the phenomenon may become more counterintuitive. If counterintuitive findings are more publishable than intuitive ones—as has classically been the case (Ledgerwood & Sherman, 2012)—the published literature may be biased in favor of findings that demonstrate splitting.

One fairly straightforward solution to this bias is that scholars can adopt an alternative mindset: Presume that any given relationship process is worth studying whether it is a function of relationship quality (i.e., the lumping explanation) or something far more specific, such as "good sex the night before" (i.e., the splitting explanation). Just as in other lumping versus splitting debates, there is rarely a principled reason to prioritize one approach a priori (Petty, Wheeler, & Bizer, 1999). Instead, we can embrace the fact that a proper description of every relational process will fall somewhere along this spectrum, and it is our job as scientists to properly depict how these processes operate, regardless of whether a global or specific construct is at play. If we collectively adopt this mindset, this form of the "isn't it just...?" critique lacks seriousness; any phenomenon we study should be worth publishing regardless of whether the mechanism is best described with an intuitive lumping versus counterintuitive splitting explanation.[1]

In conclusion, by challenging both meanings of the "isn't it just...?" critique, relationship scholars can avoid several insidious biases. First, "isn't it just...?" creates a bias toward prioritizing novelty over other important scientific goals, including replicability, cumulativeness, and generalizability. Second, "isn't it just...?" biases scholars toward counterintuitive explanations, and there is nothing inherently more scientifically valuable about counterintuitive over intuitive explanations. For these reasons, when researchers want to report a test of incremental validity, it is worth considering whether that test should be a focal point of a Results section. A

possible default Results section strategy could be that scholars first report their X–Y association of interest and then, in a subsidiary section, argue how their findings bolster (a) replicability, generalizability, cumulativeness, and a lumping explanation if X–Y is due to relationship quality or (b) bolster novelty and a splitting explanation if X–Y is not due to relationship quality. In many cases, it may be possible for scholars to persuasively make a case that either reality is empirically valuable however the tests turn out.

Of course, the "isn't it just...?" critique may be persuasive under some circumstances: Perhaps the C–Y or X–C association has been published many times in many different contexts, and the intuitive, lumping explanation connects poorly to relevant theory. In this circumstance, it may be necessary to conduct an incremental validity test. What should a researcher do?

3 | INCREMENTAL VALIDITY PROBLEM #2: YOU MIGHT NOT HAVE SEPARATE CONSTRUCTS

An important assumption in an incremental validity argument is that the focal predictor X is different from the covariate C. After all, if X and C are actually the same construct measured twice (X_1 and X_2), then there are no grounds for incremental validity testing: A focal predictor cannot (and should not) have incremental validity over itself. Thus, evidence of separate constructs as part of construct validation needs to be established as a prerequisite of incremental validity testing (Bong, 1996; Clark & Watson, 1995; Flake, Pek, & Hehman, 2017; Marsh, Craven, Hinkley, & Debus, 2003). Handling this problem requires that scholars grapple seriously with the possibility that X and C might not be separate constructs in their dataset, regardless of their chosen dependent measure (Y) (we consider the role of Y in detail in Problem #3).

3.1 | The jangle fallacy

At first glance, establishing the existence of two separate constructs seems like an easy task: For example, being romantically attracted to someone seems to be conceptually different from wanting to go on a date with him or her. Given that constructs such as these are often measured with different questionnaire items (e.g., "I am romantically attracted to _____" vs. "I would like to go on a date with _____"), it is intuitive that researchers could point to face validity as evidence of separate constructs in such cases.

Yet this line of argumentation is often insufficient. In fact, it runs the risk of committing the *jangle fallacy*—the erroneous assumption that two things are different because they have different names (Kelley, 1927). The jangle fallacy is a long-standing problem in psychology, and it impedes scientific progress by making cumulative knowledge more difficult, muddling conceptual understanding, and wasting resources (Block, 1996; Lilienfeld, Waldman, & Israel, 1994). The persistence of the jangle fallacy is readily apparent in researchers' tendency to create new labels for phenomena that have been described with other terminology (Bong, 1996; Holroyd & Coyne, 1987; Miller & Pedersen, 1999; Miller & Pollock, 1994; Rosenthal, 1994). For example, the false consensus effect has been studied under more than 15 distinct labels, such as egocentric attribution, assumed similarity, attributive projection, classical projection, and defensive projection (Miller & Pollock, 1994). The presence and negative impact of jangle fallacy is evident in many other research domains, such as personality, emotion, organizational behavior, motivation, and psychopathology (e.g., Block, 1996; Casper, Vaziri, Wayne, DeHauw, &

Greenhaus, 2018; Credé, Tynan, & Harms, 2017; Heyman & Dweck, 1992; Lilienfeld et al., 1994; Markon, 2009; Marsh et al., 2003; Weidman, Steckler, & Tracy, 2017); thus, it seems likely to be applicable to the correlated constructs often assessed by relationship researchers as well.

3.2 | Using confirmatory factor analysis to check if you have separate constructs

To avoid committing the jangle fallacy, researchers interested in making incremental validity claims need to ensure that their assumption of separate constructs X and C is supported by data. Although researchers can point to existing evidence of dissociation between two constructs in the literature (e.g., previous validity studies, meta-analyses, existing datasets), appropriate incremental validity tests (see Problem #3 below) still require that the dataset researchers use to conduct these tests contain two empirically separate constructs. If there is no evidence of separate constructs in a given dataset, researchers should not use that dataset to make incremental validity claims.

What counts as evidence of separate constructs? Researchers might be tempted to calculate the raw correlation between X and C and compare the result to some intuitive standard for what seems to be "low enough" to reflect two different constructs (e.g., $r < .60$). However, this strategy is suboptimal because low correlations between X and C might simply indicate the presence of measurement error, which typically depresses bivariate correlations (McNemar, 1946; Spearman, 1904, 1910, but see Stanley & Spence, 2014). For example, uncorrected test–retest reliability—the extent to which a construct relates to itself across measurement occasions—can reach as low as $r = .3$–$.4$ for some social psychological constructs, even when measured within 24 hr (e.g., Bar-Anan & Nosek, 2014; Dasgupta, McGhee, Greenwald, & Banaji, 2000). Therefore, even an intuitively low correlation between two variables still does not offer strong evidence that the two variables represent two separate constructs.

A second suboptimal strategy is to simply conduct the incremental validity test anyway, and if the test is successful (i.e., if X-Y is significant controlling for C), assume that the X and C variables represent two constructs. Once again, however, because constructs are imperfectly measured, random fluctuations across the two measurement occasions can cause the two variables to predict different portions of the outcome variable, even if they reflect the same construct. Thus, a purportedly successful test of incremental validity does not establish the separability of the focal predictor and covariate.

A better approach requires the researcher to formally model the relations between X and C, as well as their measurement error, and latent variable analysis—particularly confirmatory factor analysis (CFA)—is a useful tool for accomplishing this goal. CFA is an instantiation of the common factor model (Thurstone, 1947), in which the shared variability among manifest variables (e.g., variables directly measured in a study) is attributed to one or more latent variables, which are often interpreted as the underlying psychological construct(s) of research interest. Because CFA simultaneously estimates how manifest variables relate to latent variables and how latent variables relate to each other, researchers can use CFA to obtain more accurate estimates of the relations among the constructs and to model how the measured variables reflect those constructs. In the context of testing for evidence of separate constructs, CFA is particularly useful because, unlike other forms of factor analyses (e.g., exploratory factor analyses), the researcher can specify that each manifest variable in CFA loads onto only one latent variable; conceptually, this means that the researcher states that a given item in a measure reflects only

one latent construct (e.g., "I am romantically attracted to _____" is assumed to only measure one latent construct "romantic attraction" and not other latent constructs such as "dating intentions").

Because incremental validity testing assumes that the covariate and the focal predictor are separate constructs, a test for the evidence of separate constructs—even when the covariate and the focal predictor seem obviously different—should typically be conducted. In practice, evidence of separate constructs can be obtained from a comparison between two CFA models: a unifactor model in which the focal predictor X and the covariate C load onto a single latent variable and a two-factor model in which X and C load onto their respective latent variables (see Figure 1 for an example). The former models the scenario in which X and C measure a single construct; the latter models the scenario in which X and C measure two separate (but likely correlated) constructs. To obtain evidence for separate constructs, the two-factor model should provide a meaningfully better fit (e.g., a significantly lower χ^2 value based on a chi-square difference test) than the unifactor model and converge on a sensible solution (e.g., with high factor loadings in the expected directions; Bagozzi & Phillips, 1982; Joreskog, 1971).[2]

We note that the solution we propose for Problem #2 of incremental validity testing should not be taken as a general solution for the "jangle fallacy." Furthermore, a better-fitting two-factor (vs. unifactor) model does not offer definitive evidence that there must be two constructs because the better fit could be due to nonsubstantive reasons (e.g., structural similarity among one vs. another set of items). Rather, our solution should be seen as a minimum requirement for obtaining empirical evidence of X and C as separate constructs, and it sets up the use of structural equation modeling (SEM) to obtain more accurate estimates of incremental validity (see Problem #3 below).

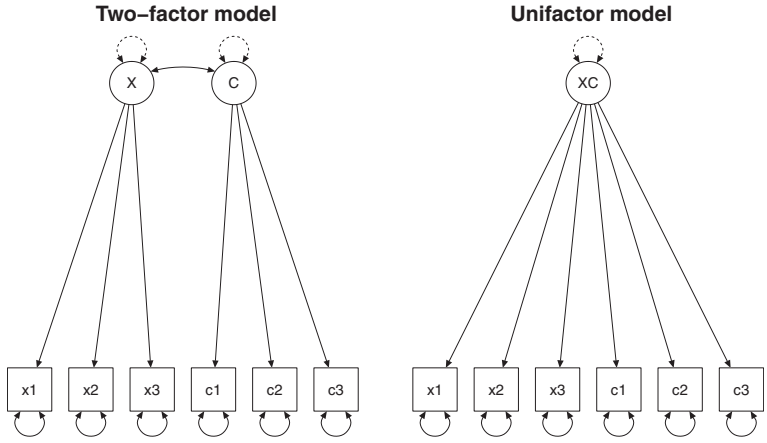

FIGURE 1 Diagram of a unifactor model and a two-factor model. In this example, the two-factor model represents a focal predictor and a covariate as two separate constructs, X and C. Each construct is conceptualized as a latent variable (represented as a circle) and measured with three items (i.e., manifest variables, represented as squares). The association between X and C is estimated (represented as a double-headed arrow). The unifactor model, on the other hand, represents a single latent construct XC measured by items in X and C (note that the unifactor model is equivalent to and can be alternatively represented as a two-factor model in which the factor correlation is set to 1). Residual variances (represented as double-headed self-loops) of manifest variables are set to be independent from each other in both models

4 | INCREMENTAL VALIDITY PROBLEM #3: YOUR RESULT MIGHT BE A TYPE I ERROR

4.1 | Pitfall of using standard multiple regressions

Once researchers ensure that the incremental validity claim they are trying to make does involve a separate focal predictor and a covariate, it becomes viable to ask if the incremental validity claim is supported by data. The most popular analysis for this purpose is a standard multiple linear regression in which the dependent variable Y is simultaneously regressed on the covariate C and the focal predictor X. If the partial regression coefficient of X is significantly different from zero, then researchers typically conclude that there is evidence for the incremental validity of X, over and above C. This approach is widespread in the close relationship literature and in myriad other social sciences, such as epidemiology, education, and econometrics (Fewell, Smith, & Sterne, 2007; Hunsley & Meyer, 2003; Shear & Zumbo, 2013; Wansbeek & Meijer, 2000). Indeed, of the 32 empirical articles published in *Personal Relationships* in 2017 that contained at least one test of incremental validity, 28 (88%) used only standard multiple regressions (or a related technique that did not account for measurement error in the covariates) for that purpose (see Table S1 of Supplemental Materials for details).

Despite its intuitive appeal, the standard multiple regression approach can result in inflated Type I error rates. The reason why Type I error rates are often inflated in this context is that the predominant approach to multiple regressions (which uses the ordinary least squares [OLS] estimator) does not account for the fact that constructs in relationship research are typically latent and, as discussed above, measured imperfectly (McNemar, 1946). To the extent that constructs are measured with error, directly entering them in a multiple regression and ignoring measurement error will result in inaccurate estimates of the true effects (Brunner & Austin, 2009; Cole & Preacher, 2014; Gustafson, 2004; Kahneman, 1965; Shear & Zumbo, 2013; Spearman, 1910).

How inaccurate might these estimates be? Recent research has shown that a moderate amount of measurement error in the focal predictor and covariate can cause Type I error rates to be alarmingly high—approaching 100% in some cases—and can even lead researchers to draw near-impossible conclusions (Culpepper & Aguinis, 2011; Shear & Zumbo, 2013; Westfall & Yarkoni, 2016). To illustrate this problem, Westfall and Yarkoni considered a classic case where a researcher documents a positive association between ice cream sales and swimming pool death rates. Even though any psychology major can immediately spot temperature as a third variable that fully explains the association, Westfall and Yarkoni illustrated how spurious evidence of the incremental validity of ice cream sales can be obtained using the multiple regression approach. Specifically, if a researcher were to measure temperature not using a thermometer but instead as subjective feelings of heat—akin to many self-reported psychological measures—they would introduce measurement error to their analysis. In simulations with this subjective measure, a multiple regression approach to incremental validity would (nearly always) lead to the conclusion that ice cream sales significantly predict swimming pool death rates over and above temperature—a (seemingly oxymoronic) highly replicable Type I error.

High Type I error rates can occur even at moderate levels of reliability (e.g., $\rho = .6–.8$; Westfall & Yarkoni, 2016). Perhaps counterintuitively, larger samples can exacerbate, rather than mitigate, the problem of high Type I error rates because larger samples detect any reliable associations with greater power, even when those associations are actually due to measurement error (this is sometimes termed "residual confounding"; Fewell et al., 2007). Furthermore, if the

dependent variable also contains measurement error, or if the incremental validity test involves more than one covariate, the Type I error rate and power will become even more challenging to estimate properly (Cole & Preacher, 2014). In short, when researchers wish to draw substantive conclusions about incremental validity, the use of multiple regression without accounting for measurement error can be misleading, even when measurement reliability seems adequate.

4.2 | Using SEM for better estimates

SEM offers an alternative analytic strategy that can remedy the pitfalls of standard multiple regression for incremental validity claims (Culpepper & Aguinis, 2011; Westfall & Yarkoni, 2016).[3] Instead of assuming that constructs are perfectly measured, SEM estimates how much latent psychological constructs contribute to the measures that represent them (vs. unrelated measurement error). This estimation process in the measurement model is the same as that of the confirmatory factor analytic approach used to check for evidence of separate constructs, as recommended above. SEM subsequently allows researchers to obtain more accurate estimates of the relations among the latent variables, as represented by the parameters in the structural component. In the context of incremental validity claims, the estimate of interest would be the path coefficient between the (latent) focal predictor X and the (latent) outcome Y, with the (latent) covariate C also included in the model. An estimate significantly different from zero can be interpreted and reported as evidence of the incremental validity of X over C on Y (see Figure 2).

Conducting a SEM analysis for incremental validity claims is typically straightforward, and most software packages that can run a CFA (as we recommend in response to problem #2 above) can run SEM, too. We provide researchers with sample R syntax using the "lavaan" package (R Core Team, 2018; Rosseel, 2012) for incremental validity testing of individual data at https://osf.io/uw7gc/ and of (distinguishable) dyadic data at https://osf.io/7tk5c/.

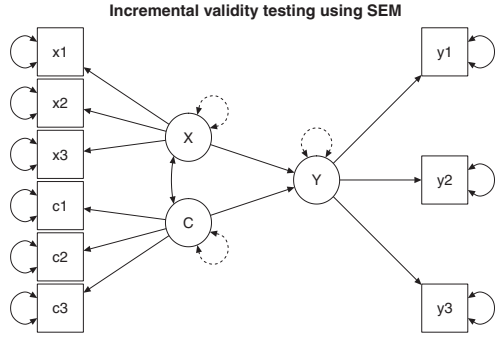

FIGURE 2 Diagram of an example structural equation model for incremental validity testing. In addition to the two-factor measurement model used to verify separation of constructs (see Figure 1), the model contains the outcome Y as a latent construct measured by manifest variables y1–y3. In this model, a significant partial regression coefficient between the focal predictor X and outcome Y (represented by the arrow pointing from X to Y) can be interpreted as evidence of the incremental validity of X over C. The outcome Y can also be modeled as a manifest variable if it is not conceptualized as a latent construct (e.g., whether a breakup has happened or not). Residual variances of manifest variables are set to be independent from each other

Nevertheless, there are a few caveats that researchers should keep in mind when taking this analytic approach.

4.2.1 | Keep an eye on statistical power

Using SEM to conduct incremental validity testing will reduce the (sometimes dramatically high) Type I error rate associated with multiple regression. However, this decision may have the cost of increasing the Type II error rate when a true incremental effect is present (Wang & Rhemtulla, 2020; Westfall & Yarkoni, 2016). This observation is somewhat counterintuitive—if SEM "accounts for" measurement error and disattenuates standardized effect sizes at the latent level, should power not increase? In reality, simulation studies suggest that the power of SEM in incremental validity testing tends to be lower than that of multiple regression because SEM appropriately calibrates the standard errors to the actual uncertainty of the data (Ledgerwood & Shrout, 2011). As a result, the standard errors of parameter estimates tend to be larger. In turn, these larger standard errors mean that the power to detect a nonzero parameter in SEM—which is the researcher's goal in an incremental validity test—is typically lower than the power to detect that same effect in a multiple regression model.

Power to detect an incremental validity path in SEM varies as a function of model characteristics, including sample size, effect size, factor loadings, and number of indicators per latent variable (Wang & Rhemtulla, 2020; see also Wolf, Harrington, Clark, & Miller, 2013). A desirable level of power (e.g., 80%) might require a much larger sample size—sometimes in the thousands—compared to a comparable multiple regression model (Westfall & Yarkoni, 2016). To obtain realistic estimates of target sample size, we recommend that researchers planning to use SEM for incremental validity testing conduct power analysis to detect their target effect of interest (i.e., the incremental effect of the focal predictor on the outcome variable). Such power analysis can be conducted using Monte Carlo simulations (Muthén & Muthén, 2002; for a user-friendly Shiny app for this type of power analysis, see pwrSEM; Wang & Rhemtulla, 2020).

4.2.2 | Consider how to estimate measurement error for single-item measures

The standard way of estimating measurement error in SEM requires multiple indicators (e.g., items in a questionnaire) per latent variable in order to estimate the shared variance of those indicators. Given that most relationship researchers typically measure their constructs with multiple items, this aspect of the procedure is relatively straightforward. However, sometimes, relationship researchers use single-item measures (e.g., in brief event-sampling questionnaires or for certain constructs such as breakup decisions). In these cases, researchers should follow a "two-step" or "single-indicator" procedure, as recommended in various SEM textbooks: First, specify (i.e., fix) an estimate or a range of estimates of the amount of measurement error in the single-item measures to the model directly; second, estimate the structural model while correcting for the fixed measurement error variances (Bollen, 1989; Hayduk, 1987; Savalei, 2019; Schumacker & Lomax, 2004). This procedure is often used because it can reduce model complexity and avoid potential issues with empirical estimation of measurement error, and it is especially beneficial in small samples (i.e., $N < 200$; Savalei, 2019). However, this procedure can be challenging because researchers often do not know the reliability of the single-item measure.

If information on its reliability is not available (e.g., in published validation work or large datasets), researchers can (a) estimate its reliability by using the Spearman-Brown formula if the item is a part of a larger inventory with known reliability information (Kuder & Richardson, 1937; Lord & Novick, 1968)[4] or (b) specify a range of plausible values of reliability and explore how sensitive results are to those specified values (see p. 14 of Westfall & Yarkoni, 2016, for an example; we direct readers interested in learning more about this procedure and its associated benefits and challenges to Oberski & Satorra, 2013, and Savalei, 2019).

4.2.3 | Interpret path coefficients in the context of model fit

Even though the path coefficients between Y and X (or C) are of chief interest here, estimates of those coefficients are only meaningful when the model adequately characterizes the data. Conversely, if the model fits the data poorly (e.g., as indicated by fit indices below "acceptable" levels, such as $CFI < 0.95$ and $RMSEA > 0.08$; Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996),[5] the path coefficients in the model should be interpreted with caution or not interpreted at all. Researchers should keep in mind how well their chosen model fit their data when interpreting specific parameters of interest and consider other theoretically sensible models as an alternative if they provide better fit. We thus reiterate our recommendation of verifying the measurement model with CFAs before interpreting the path coefficients.

5 | HOW TO TEST INCREMENTAL VALIDITY CONVINCINGLY: AN EXAMPLE

To illustrate how researchers can overcome the challenges and pitfalls associated with testing incremental validity, we turn to one recent concrete published example (Joel, Impett, Spielmann, & MacDonald, 2018). We discuss how the problems of incremental validity testing apply to the research question of this article, as well as how the authors addressed these problems in an exemplary fashion.

5.1 | Addressing incremental validity problem #1

Joel et al. (2018) tested whether participants' perceptions of their partner's dependency on a romantic relationship (operationalized as perceived partner commitment) would predict their decision to break up with the partner. The "isn't it just...?" critique in this case would state that the perception of a partner's commitment is a proxy for relationship quality and associated self-interested reasons to stay in the relationship (e.g., own relationship satisfaction and own commitment). Because it is well established on both theoretical and empirical grounds that constructs such as own satisfaction and commitment predict stay/leave decisions (Le, Dove, Agnew, Korn, & Mutso, 2010; Rusbult, 1980, 1983), meaning #1 of the "isn't it just...?" critique (i.e., the finding lacks novelty) would have been effective in this case. That is, a demonstration of novelty would arguably contribute more to the literature than an additional demonstration of replicability. Therefore, the authors needed to test the incremental validity of perceived partner commitment on stay/leave decisions.

5.2 | Addressing incremental validity problem #2

In order to argue that perceived partner commitment (i.e., focal predictor) has incremental validity over satisfaction and commitment (i.e., covariates), Joel and colleagues needed to first show that these variables represent three separate constructs. To that end, the authors first noted that perceived partner commitment did not correlate with own commitment ($r = .02$) and only moderately correlated with own satisfaction ($r = .37$; Study 2, Joel et al., 2018). Second and more convincingly, the authors assessed the relations among the three variables by modeling each as a separate latent factor. As noted above, this procedure allows researchers to account for measurement error and more accurately estimate the relations between focal predictors and covariates. The intercorrelations among the three factors varied (rs = .02–.46) and largely suggested separability of constructs. Importantly, this three-factor model fit the data adequately ($RMSEA = 0.08$, $CI_{95\%}$ [0.07, 0.08], $CFI = 0.94$, $\chi^2(114) = 441.44$, $p < .001$) and significantly better than alternative models. Specifically, the three-factor model provided a superior fit relative to a unifactor model (in which all three constructs were modeled as a single latent factor) and to several two-factor models as well. These models provide strong evidence that perceived partner commitment, own commitment, and own satisfaction were separate constructs.[6]

5.3 | Addressing incremental validity problem #3

Finally, the authors tested the incremental validity of perceived partner commitment by regressing breakup decisions (i.e., outcome variable) on all three constructs using SEM. A path coefficient from perceived partner commitment to breakup decision that is significantly different from zero would suggest that perceived partner commitment predicted breakup over and above the other two constructs, and the model indeed detected a significant coefficient in the predicted direction, $b = -0.04$, $SE = 0.02$, $p = .02$. The authors were therefore justified in drawing an incremental validity conclusion: Perceived partner commitment to a romantic relationship predicts stay/leave decisions over and above self-interested reasons to stay in the relationship (i.e., own satisfaction and own commitment).

6 | INCREMENTAL VALIDITY TESTING OF DYADIC AND LONGITUDINAL DATA

We have focused our discussion on incremental validity testing of individual data in cross-sectional study designs, but of course, relationship researchers often collect dyadic and longitudinal data. The incremental validity issues we discussed above have parallels in these designs: For example, a researcher with a dataset of heterosexual married couples may wish to show that the man's affect predicts outcomes over and above the woman's affect (e.g., in an Actor–Partner Interdependence Model; Kashy & Kenny, 2000) or show in a longitudinal study that subjective stress at Time 1 predicts Time 2 outcomes over and above Time 2 stress (e.g., in a cross-lagged panel model). Problems #1 and #2 discussed above rarely apply in these contexts: Presumably, a researcher would want to test a dyadic or longitudinal question because the equivalent question at the individual level has a well-established answer, and therefore, testing dyadic or longitudinal questions adds novelty. Similarly, constructs reported by two people or by one person at two different time points are not "the same thing," as long as the definition of the construct

permits it to vary by individuals within a dyad or across time within an individual. However, problem #3 nearly always still applies: Type I error in estimating incremental effects from these types of datasets can be inflated by measurement error for the same reasons that we have detailed above.

In fact, problem #3 is potentially even more challenging to address in the dyadic and longitudinal cases because of recently identified issues related to variance decomposition failures (Hamaker, Kuiper, & Grasman, 2015). The incremental validity claims researchers wish to make from dyadic and longitudinal data are often at Level 1 (i.e., the within-dyad level for dyadic data, the within-person level for longitudinal data) rather than Level 2. Given the hierarchical structure of data (e.g., individuals nested within dyads, or measurement occasions nested within individuals), incremental validity testing with dyadic and longitudinal research questions needs to correctly partial out Level 2 variances; otherwise, the incremental validity paths are forced to absorb any stable variance at the higher level (e.g., stable characteristics of the dyad or of the individual), which inflates Type I error (for detailed discussions of this point, see Hamaker et al., 2015; Rogosa, 1980; Schuurman & Hamaker, 2019).

Recent developments in statistical techniques that combine the strengths of multilevel modeling and SEM allow researchers to address challenges that arise from both measurement error and variance decomposition. For dyadic data, researchers can model their variables as latent variables in their dyadic analysis models, such as Actor-Partner Interdependence Models (APIM), in order to improve estimates; this is a key advantage to conducting APIM in the SEM framework (Ledermann & Kenny, 2017; Orth, 2013; Stas, Kenny, Mayer, & Loeys, 2018). Alternatively, researchers can also directly fix a certain level of unreliability in their variables, similar to the "two-step" procedure we recommend above for single-indicator variables. For the latter approach, we direct interested readers to APIM_SEM, a free, user-friendly Shiny web application (http://lavaan.org/APIM_SEM/; Stas et al., 2018) that provides researchers with the option of correcting for unreliability. For longitudinal data, random intercepts cross-lagged panel models (RI-CLPM; Hamaker et al., 2015), a type of SEM that models the multilevel nature of longitudinal data, allows researchers to obtain more accurate estimates from incremental validity testing.

7 | CONCLUSION

Incremental validity testing is a common practice in relationship science, but there are reasons to be cautious about both why and how it is done. We discussed three potential problems with incremental validity testing and proposed solutions. Researchers seeking to demonstrate the incremental validity of their predictor should consider (a) whether their findings are truly uninteresting and/or well established without evidence of it, (b) demonstrating that their predictor represents a separate construct from their covariate, and (c) conducting SEM to avoid inflating Type I error. We provided guidelines on how to address these questions and illustrated our solutions in a recently published example in Joel et al. (2018), and we noted how the Type I error inflation issues continue to apply in the common cases where relationship scholars ask dyadic and longitudinal research questions.

Across the sciences, scholars have recently been revisiting and improving their methodological and statistical practices to ensure that their conclusions are robust and replicable (Spellman, 2015). Improving incremental validity testing should be a core focus of this broader movement: After all, incremental validity tests are often conducted in the service of novelty, and in general,

an overemphasis on novelty can impede the creation of cumulative and replicable scientific literatures (Giner-Sorolla, 2012). In addition, incremental validity testing involves a tradeoff between Type I error and Type II error (Finkel et al., 2017; LeBel et al., 2017), although in this case, the Type I error risk is remarkably high—considerably higher than risks associated with other recently highlighted practices (e.g., optional stopping in data collection, using independent covariates in experimental designs; Sagarin, Ambler, & Lee, 2014; Simmons, Nelson, & Simonsohn, 2011; Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017). However, some commonly discussed solutions to methodological problems in psychology will not help resolve problems with incremental validity. For example, preregistered analysis plans and larger sample sizes can help reduce Type I error in many corners of psychological science (Ledgerwood, 2019). However, if a scholar uses multiple regression to conduct incremental validity tests, preregistered analysis plans will not mitigate Type I error, and larger *N* will actually increase it (Westfall & Yarkoni, 2016). We hope that researchers (as well as editors and reviewers) will keep these considerations in mind as they approach issues surrounding incremental validity and engage in practices that would help build more robust, informative evidence in the service of relationship science.

ACKNOWLEDGMENTS

We thank Eli Finkel and Alison Ledgerwood for their helpful comments on an earlier version of this article, Samantha Joel for sharing original data for reanalysis, and Maeve Kelly for assistance with article coding.

ORCID

Y. Andre Wang https://orcid.org/0000-0002-5729-7373

ENDNOTES

[1] In principle, it is possible for the lumping solution to be counterintuitive and the splitting solution to be intuitive, but in our experience, this is uncommon.

[2] Although it is possible in principle for the researcher to examine how well the unifactor or the two-factor model alone fits the data, we caution against such an approach because of the limited utility of relying on global fit indices from any given model (e.g., Hopwood & Donnellan, 2010; McDonald & Ho, 2002; Sobel & Bohrnstedt, 1985; Williams & O'Boyle Jr, 2011). For example, chi-square test of exact fit, an index frequently used to assess model fit, tends to reject models that are "trivially misspecified" in large samples (Bentler & Bonett, 1980). This means that researchers might interpret a poor exact fit of a unifactor model as evidence of separate constructs when it is more accurately interpreted as a methodological artifact.

[3] Other methods for correcting measurement error when controlling for covariates exist, such as errors-in-variables models (Warren, White, & Fuller, 1974), Lord's (1960) method, and Raaijmakers and Pieters's (1987) R&P method. For a summary and comparison of these methods, see Culpepper and Aguinis (2011). We focus on SEM here because it is effective against Type I error rate inflation, familiar to relationship researchers, and is more flexible than other methods (e.g., does not assume equal error variances in the predictor and the outcome DV; Raaijmakers & Pieters, 1987).

[4] As a primer of intuition, if a six-item scale has an overall reliability of .8 and the items are parallel (i.e., equal item reliability), then the expected reliability of a single item is .4. This can be shown mathematically by using the Spearman-Brown formula

$$r^{*}_{xx'} = \frac{kr_{xx'}}{1+(k-1)r_{xx'}},$$

where $r^{*}_{xx'}$ is the expected overall reliability of the scale, k represents test length (in the case of calculating expected reliability of multiple-item scales, the number of items), and $r_{xx'}$ is the expected reliability of a single item. Given overall scale reliability, rearranging the terms gives us the expected item reliability as

$r_{xx'} = \frac{r^*_{xx'}}{k + (1-k) r^*_{xx'}}$.

Given $r^*_{xx'} = .8$ and $k = 6$,

$r_{xx'} = \frac{.8}{6 + (1-6) \times .8} = .4$.

[5] Note, however, that these recommendations should be considered in the context of the specific models researchers have; they should not be seen as "cutoffs" (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011).

[6] Note that the model comparisons reported in Joel et al. (2018) were conducted with the outcome variable included in the models. Given that testing for separability of constructs is a measurement question, a more direct test would only compare the measurement models through CFAs (i.e., models that only include the three latent factors and their indicators and allow intercorrelations among latent factors to be freely estimated), without the outcome variable regressed on the latent factors (see Anderson & Gerbing, 1988). Additional analyses (using the data provided by Dr. Joel; S. Joel, personal communication, October 3, 2018) that directly compared among the measurement models suggested that the three-factor model still provided the best data fit, consistent with the conclusions in this article.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly*, *27*, 459–489.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*, 668–688.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Block, J. (1996). Some jangly remarks on Baumeister and Heatherton. *Psychological Inquiry*, *7*, 28–32.

Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bong, M. (1996). Problems in academic motivation research and advantages and disadvantages of their solutions. *Contemporary Educational Psychology*, *21*, 149–165.

Brunner, L. J., & Austin, P. C. (2009). Inflation of type I error rate in multiple regression when independent variables are measured with error. *Canadian Journal of Statistics*, *37*, 33–46.

Campbell, L., Loving, T. J., & LeBel, E. P. (2014). Enhancing transparency of the research process to increase accuracy of findings: A guide for relationship researchers. *Personal Relationships*, *21*, 531–545.

Campbell, L., Simpson, J. A., Boldry, J. G., & Rubin, H. (2010). Trust, variability in relationship evaluations, and relationship processes. *Journal of Personality and Social Psychology*, *99*, 14–31.

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.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., ... Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*, 750–764.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300–315.

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*, 492–511.

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, *16*, 166–178.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for white Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*, 316–328.

Eastwick, P. W., & Finkel, E. J. (2008). The attachment system in fledging relationships: An activating role for attachment anxiety. *Journal of Personality and Social Psychology*, *95*, 628–647.

Eastwick, P. W., & Finkel, E. J. (2012). The evolutionary armistice: Attachment bonds moderate the function of ovulatory cycle adaptations. *Personality and Social Psychology Bulletin*, *38*, 174–184.

Fewell, Z., Smith, G. D., & Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology*, *166*, 646–655.

Fincham, F. D., Garnier, P. C., Gano-Phillips, S., & Osborne, L. N. (1995). Preinteraction expectations, marital satisfaction, and accessibility: A new look at sentiment override. *Journal of Family Psychology*, *9*, 3–14.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*, 244–253.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378.

Fletcher, G. J. O., Simpson, J. A., & Thomas, G. (2000). The measurement of perceived relationship quality components: A confirmatory factor analytic approach. *Personality and Social Psychology Bulletin*, *26*, 340–354.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571.

Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Boca Raton, FL: Chapman & Hall/CRC.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102–116.

Hayduk, L. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins University Press.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*, 319–336.

Heyman, G. D., & Dweck, C. S. (1992). Achievement goals and intrinsic motivation: Their relation and their role in adaptive motivation. *Motivation and Emotion*, *16*, 231–247.

Holroyd, K. A., & Coyne, J. (1987). Personality and health in the 1980s: Psychosomatic medicine revisited? *Journal of Personality*, *55*, 359–375.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446–455.

Joel, S., Impett, E. A., Spielmann, S. S., & MacDonald, G. (2018). How interdependent are stay/leave decisions? On staying in the relationship for the sake of the romantic partner. *Journal of Personality and Social Psychology*, *115*, 805–824.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Kahneman, D. (1965). Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin*, *64*, 326–329.

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 451–477). New York, NY: Cambridge University Press.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York, NY: World Book.

Kruglanski, A. W., Chernikova, M., & Jasko, K. (2016). Aspects of motivation: Reflections on Roy Baumeister's essay. *Motivation and Emotion*, *40*, 11–15.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Le, B., & Agnew, C. R. (2003). Commitment and its theorized determinants: A meta-analysis of the investment model. *Personal Relationships*, *10*, 37–57.

Le, B., Dove, N. L., Agnew, C. R., Korn, M. S., & Mutso, A. A. (2010). Predicting nonmarital romantic relationship dissolution: A meta-analytic synthesis. *Personal Relationships*, *17*, 377–390.

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*, 230–243.

Ledermann, T., & Kenny, D. A. (2017). Analyzing dyadic data with multilevel modeling versus structural equation modeling: A tale of two methods. *Journal of Family Psychology*, *31*, 442–452.

Ledgerwood, A. (2014). Introduction to the special section on moving toward a cumulative science: Maximizing what our research can tell us. *Perspectives on Psychological Science*, *9*, 610–611.

Ledgerwood, A. (2019). New developments in research methods. In E. J. Finkel & R. F. Baumeister (Eds.), *Advanced social psychology* (pp. 39–61). New York, NY: Oxford University Press.

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, *7*, 60–66.

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*, 1174–1188.

Lilienfeld, S. O., Waldman, I. D., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice*, *1*, 71–83.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*, 307–321.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising*, *24*, 77–87.

Markon, K. E. (2009). Hierarchies in the structure of personality traits. *Social and Personality Psychology Compass*, *3*, 812–826.

Marsh, H. W., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the big-two-factor theory of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research*, *38*, 189–224.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289–374.

Miller, N., & Pedersen, W. C. (1999). Assessing process distinctiveness. *Psychological Inquiry*, *10*, 150–156.

Miller, N., & Pollock, V. E. (1994). Meta-analysis and some science-compromising problems in social psychology. In W. R. Shadish & S. Fuller (Eds.), *The social psychology of science* (pp. 230–261). New York, NY: Guilford.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, *5*, 85–90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality*, *8*, 21–29.

Oberski, D. L., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling*, *20*, 409–428.

Orth, U. (2013). How large are actor and partner effects of personality on relationship satisfaction? The importance of controlling for shared method variance. *Personality and Social Psychology Bulletin*, *39*, 1359–1372.

Petty, R. E., Wheeler, S. C., & Bizer, G. Y. (1999). Is there one persuasion process or more? Lumping versus splitting in attitude change theories. *Psychological Inquiry*, *10*, 156–163.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from. https://www.R-project.org/

Raaijmakers, J. G. W., & Pieters, L. P. M. (1987). Measurement error and ANCOVA: Functional and structural relationship approaches. *Psychometrika*, *52*, 521–538.

Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*, 245–258.

Rosenthal, R. (1994). On being one's own case study: Experimenter effects in behavioral research-30 years later. In W. R. Shadish & S. Fuller (Eds.), *The social psychology of science* (pp. 214–229). New York, NY: Guilford.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rusbult, C. E. (1980). Commitment and satisfaction in romantic associations: A test of the investment model. *Journal of Experimental Social Psychology*, *16*, 172–186.

Rusbult, C. E. (1983). A longitudinal test of the investment model: The development (and deterioration) of satisfaction and commitment in heterosexual involvements. *Journal of Personality and Social Psychology*, *45*, 101–117.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293–304.

Savalei, V. (2019). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods*, *24*, 352–370.

Schumacker, R., & Lomax, R. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, *24*, 70–91.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, *23*, 153–158.

Shear, B. R., & Zumbo, B. D. (2013). False positives in multiple regression: Unanticipated consequences of measurement error in the predictor variables. *Educational and Psychological Measurement*, *73*, 733–756.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. *Sociological Methodology*, *15*, 152–178.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*, 886–899.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318.

Stas, L., Kenny, D. A., Mayer, A., & Loeys, T. (2018). Giving dyadic data analysis away: A user-friendly app for actor–partner interdependence models. *Personal Relationships*, *25*, 103–119.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Wang, Y. A., & Rhemtulla, M. (2020). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. Davis, CA: University of California. Unpublished manuscript.

Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and type I error inflation. *Journal of Experimental Social Psychology*, *72*, 118–124.

Wansbeek, T., & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Amsterdam, The Netherlands: Elsevier.

Warren, R. D., White, J. K., & Fuller, W. A. (1974). An errors-in-variables analysis of managerial role performance. *Journal of the American Statistical Association*, *69*, 886–893.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*, 267–295.

Weiss, R. L. (1980). Strategic behavioral marital therapy: Toward a model for assessment and intervention. In J. P. Vincent (Ed.), *Advances in family intervention, assessment and theory* (Vol. 1, pp. 229–271). Greenwich, CT: JAI Press.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS One*, *11*, e0152719.

Williams, L. J., & O'Boyle, E., Jr. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods*, *14*, 350–369.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*, 913–934.

SUPPORTING INFORMATION

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

How to cite this article: Wang YA, Eastwick PW. Solutions to the problems of incremental validity testing in relationship science. *Pers Relationship*. 2020;27:156–175. https://doi.org/10.1111/pere.12309