Organizational Research Methods
15(1) 17-32

DOI: 10.1177/1094428110392383
http://orm.sagepub.com
SAGE

Understanding the Impact of Convergent Validity on Research Results

Kevin D. Carlson[1] and Andrew O. Herdman[2]

Abstract

Using different measures of constructs in research to develop robust evidence of relationships and effects is seen as good methodological practice. This assumes these measures possess high convergent validity. However, proxies—alternative measures of the same construct—are rarely perfectly convergent. Although some convergence is preferred to none, this study demonstrates that even modest departures from perfect convergent validity can result in substantial differences in the magnitudes of findings, creating challenges for the accumulation and interpretation of research. Using data from published research, the authors find that substantial differences in findings between studies using desired and proxy variables occur even at levels of convergent validity as high as $r = .85$. Implications of using measures with less-than-ideal convergent validity for the interpretability of research results are examined. Convergent validities above $r = .70$ are recommended, whereas those below $r = .50$ should be avoided. Researchers are encouraged to develop and report convergent validity data.

Keywords

field research methods, research design, measurement design, content validity, reliability and validity, construct validation procedures

Using a variety of research methods to demonstrate the robustness of a study's findings is generally viewed as a hallmark of good research practice. For this reason, tests of the robustness of findings are common in organizational research (e.g., Earle, Spicer, & Peter, 2010; Guler & Guillen, 2010; McDonald & Westphal, 2010). The use of different research protocols, samples, measures, and statistical analyses are all intended to provide assurance that findings are not idiosyncratic to components of methodology. For example, in discussing opportunity's role in decisions to start new businesses, Short, Ketchen, Shook, and Ireland (2010) argue that "researchers should use multiple measures of opportunity to achieve triangulation and enhance confidence that the concept is being

[1] Department of Management, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2] Department of Management, East Carolina University, Greenville, NC, USA

Corresponding Author:
Kevin D. Carlson, Department of Management, Virginia Polytechnic Institute and State University, 2094 Pamplin Hall, Blacksburg, VA 24061
Email: kevinc@vt.edu

captured" (p. 53). However, this practice assumes that alternative measures capture, or very closely approximate, the same construct.

Convergent validity reflects the extent to which two measures capture a common construct. Alternative measures that provide less-than-perfect convergent validity introduce ambiguities that interfere with the development of meaningful interpretations of findings within and across studies. Early recommendations to researchers on appropriate levels of convergent validity offered only broad guidelines. Campbell and Fiske (1959), for instance, argued that "at the current stage of psychological progress, the crucial requirement is the demonstration of some convergence, not complete congruence" (p. 102). They reviewed data from 12 studies reporting 144 estimates of convergent validities ranging from $r = .02$ to $r = .82$. Some convergence in this case appears to be any level of convergent validity different from zero. Similarly, Nunnally (1967) simply suggests that high convergence will be represented by "highly correlated" measures, whereas measures that are correlated "near zero" suggest weak or no convergence (p. 91). However, the level of association necessary to be "highly" correlated is undefined. Convergent validity of $r = .00$ would clearly indicate no convergence. Conversely, convergent validities of $r = 1.0$ suggest perfect convergence. Most convergent validities in real research reside between these extremes. Without more specific guidance, researchers reach logically inconsistent conclusions, arguing that a convergent validity as low as $r = .28$ (e.g., Larraza-Kintana, Wiseman, Gomez-Mejia, & Welbourne, 2007) indicates that measures converge, whereas a convergent validity as high as $r = .75$ (e.g., Podsakoff, Whiting, Podsakoff, & Blume, 2009) signals that measures are distinct.

Research evidence suggests that actual levels of convergent validity in organizational research still vary widely. Although the construct validity of measures, and the convergent validity of alternatives, is expected to be less well developed early in the study of a phenomena (Nunnally, 1967), poor convergent validity at any point will affect the magnitudes and interpretability of research findings. What is currently not well understood is how "some convergence" affects results. These effects are difficult to identify because they are not readily observed in individual studies and only become apparent in reviews of findings across studies. In the current study, we develop estimates of the impact on findings caused by declining levels of convergent validities between alternative measures (e.g., Chen & Miller, 2007; Lee, 2007). Understanding these impacts is important for enhancing confidence in the results of individual studies and for the interpretations of research literatures.

Measurement and Convergent Validity

Choosing among alternative measures of the same construct—sometimes called *proxies*—is an aspect of measurement. In classical measurement theory (Novick, 1966), this process begins with a specific research question and one or more focal theories that determine the specific constructs and relationships to be examined. In Figure 1, theory predicts that two constructs, χ and ψ, are associated in the manner represented by ρ. In this depiction, X is a measure selected to represent χ. Y is a measure selected to represent ψ. Data for measures X and Y are collected, and the properties of these observed measures and the association between them (i.e., r) are used to infer the properties of the unobservable constructs.

In theory testing, inferring properties of unobservable constructs from data for measures assumes that those measures accurately reflect the intended constructs. Data that are unexpected or inconsistent with existing theory challenge researchers to extend or reformulate theories in ways that can account for both new and existing data. Strong correspondence of measures to constructs supports meaningful inferences from observed data to the properties of theories and, ultimately, their correspondence to real-world phenomena. Strong correspondence makes it easier to identify and reject

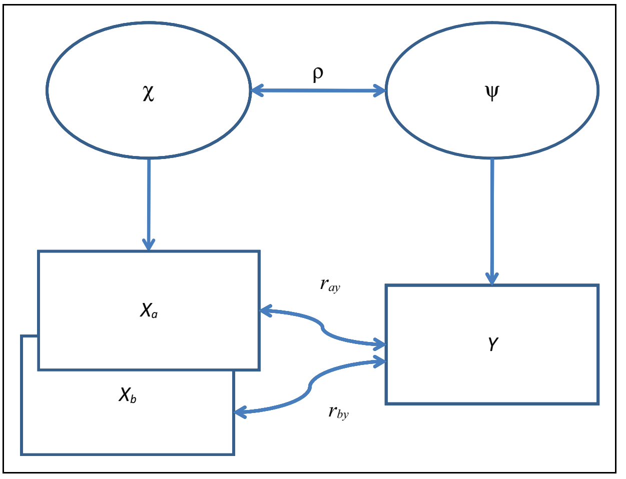

Figure 1. Visual depiction of the relationship between constructs (χ and ψ) and measures (X and Y) selected to represent those constructs in research

Note: The concepts of reliability and validity refer to the relationships between constructs and measures (i.e., χ with X_a or X_b or ψ with Y). Substitution decisions refer to the relationships between proxies (i.e., X_a vs. X_b). The observed relationship r between measures is used to make inferences regarding the properties of the unobserved (theorized) relationship ρ between constructs.

weak or incorrect theories and to build support for those that offer superior explanations. Weak correspondence provides less certainty that data actually reflect the properties of the intended constructs. When measures do not correspond to constructs, ambiguity is introduced into the meaning of the observed data. As correspondence drops, so does the value of a measure for both theory testing and theory building.

Selecting measures based on the properties of unobservable constructs is consistent with classical measurement theory's (Novick, 1966) philosophy of reflective measurement. In reflective measurement, the definition of the construct offered in theory guides the selection or development of measures. Two properties of the relationships between constructs and measures (e.g., χ with X in Figure 1)—construct validity and reliability—are discussed in classical measurement theory. Construct validity is a property of a measure that describes how well the measure captures the properties of the construct. It is best viewed as a continuum with measures possessing degrees of construct validity, rather than a determination that a measure simply is or is not construct valid.

Reliability is also a property of the measure and refers to the extent to which the measure minimizes errors of measurement and yields consistent numeric scores across assessment events. Construct validity is the more important, but more difficult, measurement property to evaluate. No single test of the construct validity of measures exists (Schwab, 1980). Instead, researchers rely on various forms of evidence to determine the extent to which a measure possesses the properties expected of the focal construct (Nunnally, 1967). Unlike reliability, for which several approaches for developing numerical estimates exist (Traub, 1994), no means of developing numerical estimates of construct validity currently exist.

In contrast to construct validity's focus on the construct–measure relationship, convergent validity refers to the relationship between measures (i.e., X_a vs. X_b in Figure 1). Thus, convergent validity does not address the construct validity of either measure directly (i.e., X_a or X_b vs. χ). Instead, it reflects the extent to which two measures capture the same information. The more similar the information they capture, the more likely they are to produce equivalent research results. Although convergent validity is not equivalent to construct validity, it is a form of evidence used to judge the construct validity of a measure. This evidence of convergent validity is commonly assessed using the magnitude of the zero-order correlation between the proxy and another closely related measure (e.g., Bakker, Demerouti, & Dollard, 2008; Guthrie, 2001; Kale & Singh, 2007; Toh, Morgeson, & Campion, 2008). If two measures are hypothesized to represent the same construct, a strong correlation between these measures suggests that the measures capture the construct. Correlations closer to $r = 1.0$ indicate stronger convergent validity. However, although strong convergent validity is a necessary condition for construct validity, it is not sufficient. As Nunnally (1967) notes, ‘‘Both tests may measure the same wrong things’’ (pp. 82-83). Nevertheless, a lack of convergent validity argues against construct validity. As the correlation between two measures declines, the likelihood that they capture the same construct also falls. When convergent validity is weak, one or both variables do not capture the intended construct well. Consequently, evidence of weak convergent validity introduces ambiguity into the meaning of research results.

Use of Proxies in Research Design

The use of proxies occurs in nearly all research domains and for a variety of reasons. Beyond their role in developing robust evidence, the use of these alternative measures occurs when desired measures are not available (e.g., Acquaah, 2007; Bercovitz & Mitchell, 2007); when they are known to be inaccurate (Acquaah, 2007; Gong, Law, Chang, & Xin, 2009); or when organizational policy, privacy concerns, or an unwillingness to release sensitive competitive information limits access to data for the ideal measures of key constructs (e.g., Boyd, Gove, & Hitt, 2005; Scandura & Williams, 2000). For example, measuring individual performance presents choices between performance records (e.g., units produced or error rates from production records or sales volume), self-report measures, or supervisor ratings. In personality research, peer and supervisor ratings are examined as alternatives to traditional self-report assessments (e.g., Funder & West, 1993). In person–organization fit, researchers choose among measures developed using difference scores, profile indices, or direct assessments (Kristof-Brown, Zimmerman, & Johnson, 2005). Likewise, a variety of proxies are found in research on organizational performance (Combs, Crook, & Shook, 2005), human resource configurations (Toh et al. 2008), knowledge transfer (Fey & Furu, 2008), and social capital (Acquaah, 2007).

A critical issue for researchers is judging when the convergent validity of proxies is strong enough to permit one measure to substitute for another in a research design. A strong correlation between proxies suggests convergent validity—the two measures capture the same information and will act the same when used in research designs. Two perfectly convergent measures will produce numerically equal correlations with other study variables and equivalent regression coefficients in multiple regression analyses. Using a proxy that is a perfect substitute, therefore, will not affect a study’s findings.

Weaker convergent validity, signaled by deviations from $r = 1.0$, highlight problems in the correspondence of measure to construct. Indeed, most proxies are imperfect substitutes. Clearly, perfectly independent proxies ($r = .00$) do not converge and therefore cannot be substituted. These proxies will produce relationships with other variables that cannot be determined a priori. Conversely, uncorrelated measures can produce identical relationships with a third variable by chance, but this outcome is highly improbable. More likely, the use of nonconverging proxies will result in different findings and conclusions.

Between perfect substitutes and perfect independence is a range of values reflecting less-than-perfect convergence but stronger convergence than that found in unrelated measures. Discussions of the substitutability of proxies with convergent validities in this range reach varied conclusions. Bommer, Johnson, Rich, Podsakoff, and MacKenzie (1995) meta-analyzed data from 65 studies that employed subjective and objective measures of performance. They found an average corrected correlation between subjective and objective measures of $\rho = .39$ ($SD_\rho = .12$). Similarly, Wall et al. (2004) reported relationships between subjective and objective measures of organization performance ranging from $r = .26$ to $r = .65$. In both studies, the authors cautioned against substitution at these levels of convergent validity but failed to offer specific guidance on levels sufficient to justify substitutions. Murphy and DeShon (2000) suggested that interrater correlations of ratings of individual performance greater than .50 are rare. Nevertheless, ratings from different sources are used as substitutes in attempts to avoid concerns over common method variance from same-source data (e.g., Bakker et al., 2008; Bommer, Dierdorff, & Rubin, 2007).

At the same time, correlations ranging in magnitude from $r = .45$ to $r = .75$ are offered as evidence of the relative distinctiveness of various person–organization fit constructs (Kristof-Brown et al., 2005). Podsakoff et al. (2009) argued that meta-analytic estimates of corrected correlations of $r_c = .75$ (equivalent to observed correlations of approximately $r = .64$) indicate that measures of organizational citizenship behavior–individual and organizational citizenship behavior–organizational were distinguishable from each other. Obviously, the stronger the association between two measures, the greater the chance these two measures capture the same construct and will produce equivalent empirical findings. As convergent validity weakens, though, more ambiguity around construct validity is introduced, and it becomes more likely that empirical results for studies using substitutes will differ depending on the substitute measure selected.

Convergent Validity's Impact on Research Findings

Evidence that weak convergent validity has an impact on research findings is found in meta-analytic reviews of research. Frequently in these analyses, method or measurement approach is identified as a moderator of findings. A troublesome characteristic of studies that use proxies is that these research designs typically do not include more than one measure representing each construct (Boyd et al., 2005). For this reason, the consequences of weak convergent validity are not evident in the results of individual studies. Two studies that employ different proxies with less-than-perfect convergence will likely produce different effect sizes. As a result, tests of the homogeneity of effect sizes are likely to indicate that studies vary to a greater extent than can be accounted for simply by sampling error (Hunter & Schmidt, 2004). When variability in effect sizes across studies exceeds levels attributable to sampling error, the relationship between constructs cannot be depicted as a single value. Instead, a range of potential effect sizes exists. The greater the unexplained variance, the wider the range of potential values and the more difficult it becomes to meaningfully interpret the meta-analytic findings. Consequently, in these instances, meta-analysts typically search for moderators that account for the extra variation in results across studies.

Bowling, Hendricks, and Wagner (2008), for example, review research on the relationship of positive affectivity with measures of facets of satisfaction. They found positive affectivity to be related to the work itself ($\rho = .31$, $SD_\rho = .20$). The large value of SD_ρ suggests the presence of moderators. In conducting the study, the authors observed that some studies used measures of extraversion rather than a specific measure of positive affectivity. Watson, Wiese, Vaidya, and Tellegen (1999) report a convergent validity between measures of positive affectivity and extraversion of $r = .51$. Subgrouping studies by positive affectivity measure, Bowling et al. found substantial differences in findings. Measures of extraversion demonstrated much weaker associations with the work itself ($\rho = .11$, $SD_\rho = .14$) than did measures of positive affectivity ($\rho = .52$, $SD_\rho = .13$). The challenge for researchers is determining whether $\rho = .31$, $\rho = .11$, or $\rho = .51$ best captures the relationship between positive affectivity and satisfaction with the work itself.

A more dramatic example is offered in a series of articles examining the risk propensity of entrepreneurs by Stewart and Roth (2001, 2004) and Miner and Raju (2004). Stewart and Roth (2001) report the results of a meta-analysis indicating entrepreneurs are more risk seeking than managers

($d = .31$, $SD_d = .00$). Miner and Raju identify 14 additional studies not included in Stewart and Roth's (2001) analysis relevant to the relationship that reaches the opposite conclusion—entrepreneurs are less risk seeking ($d = –.43$, $SD_d = .316$). In comparing the studies, the authors focus on different measures of risk propensity. Studies in Stewart and Roth's (2001) analysis primarily use the Risk Taking scale of the Jackson Personality Inventory (Jackson, 1976), which captures self-attributive motives, whereas Miner and Raju's analysis centers on risk taking measured using the Risk Avoidance Scale of the Miner Sentence Completion Scale–Form T, a projective assessment where higher scores indicate risk avoidance. Spangler (1992) reports estimates of convergent validity ranging from $r = .10$ to $r = .15$. As Miner and Raju's findings demonstrate, the reader is left to choose between three estimates of the risk propensity of entrepreneurs. Are they risk avoidant ($d = –.43$), risk seeking ($d = .31$), or somewhere in between ($d = .12$, from the combined results)?

Conversely, stronger convergent validities appear to be associated with more consistent results. In a study of the relationship of self-efficacy to work-related performance, Judge, Jackson, Shaw, Scott, and Rich (2007) report the overall relationship as $\rho = .37$. However, two approaches to assessing self-efficacy, grid and Likert-type measures, are used. Maurer and Andrews (2000) and Maurer and Pierce (1998) report convergent validities for these assessment approaches ranging from $r = .69$ to $r = .77$. Moderator analyses conducted by Judge et al. show that grid-based measures of self-efficacy produce stronger associations with work-related performance ($\rho = .44$, $SD_\rho = .24$) than do Likert-based measures ($\rho = .32$, $SD_\rho = .19$), but only slightly. However, the substantial variability in these estimates precludes unambiguous interpretations of the magnitudes of these relationships.

Gauging Substitutability

A critical issue for researchers is to understand the impact of the nonconvergence of proxies on research results. An important factor in decisions to use proxies is to understand when and to what extent less-than-perfect convergent validities begin to affect research findings and, in turn, the interpretation of results. For correlational research, a meaningful difference in findings would be one that alters our understanding of the focal relationship. For example, although zero-order correlations are generally reported to two decimals, it would be exceedingly difficult to make a practical distinction between a correlation of $r = .19$ and $r = .20$. Bivariate plots of these relationships are virtually indistinguishable. In fact, it may take differences as large as $r = .10$ or more for decision makers' interpretations of levels of association to change or, more important, for that change to affect related decisions. Bedeian, Sturman, and Streiner (2009) reach the same conclusion in their examination of the reporting and interpretation of correlations. Although it is generally understood that diminishing levels of convergence will result in a higher probability of divergent findings, the extent of these effects for different levels of convergent validity is unknown.

The Potential for Inconsistent Findings

Although some convergent validity is better than no convergent validity, the fundamental question before researchers is determining how much convergent validity is enough. An alternative is to ask how much the relationship between a proxy and an outcome (r_{by}) could differ from the magnitude of the relationship of the desired measure with the same outcome (r_{ay}) at various levels of convergent validity. Here, r_{ab} is an estimate of the convergent validity of the proxy measure X_b as related to the desired variable X_a.

Methods for estimating the potential value of an unknown correlation given the values of two other correlations in a three-variable system are examined in efforts to generate possible positive definite 3×3 correlation matrixes (Glass & Collins, 1970; Stanley & Wang, 1969). Budden,

Table 1. Values of Minimum and Maximum Possible Values of r_{by}, Given Values for r_{ay} and r_{ab}

	r_{ay}								
r_{ab}	0	.30	.50	.70	.80	.85	.90	.95	1.00
0	−1.00, 1.00	−.95, .95	−.87, .87	−.71, .71	−.60, .60	−.53, .53	−.44, .44	−.31. .31	.00, .00
	1.00	1.25	1.37	1.41	1.40 *(.20)*	1.38 *(.32)*	1.34 *(.56)*	1.26 *(.64)*	1.00 *(1.00)*
.30	−.95, .95	−.82, 1.00	−.68, .98	−.47, .89	−.33, .81	−.25, .76	−.15, .69	−.01, .58	.3, .3
	.95	1.12	1.18	1.17	1.13	1.10 (.09)	1.05 (.31)	.96 (.37)	.70 *(.70)*
.50	−.87, .87	−.68, .98	−.50, 1.00	−.27, .97	−.12, .92	−.03, .88	.07, .83	.21, .75	.50, .50
	.87	.98	1.00	.97	.92	.88	.83 (.07)	.74 (.20)	.50 *(.50)*
.70	−.71, .71	−.47, .89	−.27, .97	−.02, 1.00	.13, .99	.22, .97	.32, .94	.44, .89	.70, .70
	.71	.77	.77	.72	.67	.63	.58	.51 (.06)	.30 *(.30)*
.80	−.60, .60	−.33, .81	−.12, .92	.13, .99	.28, 1.00	.36, .99	.46, .98	.57, .95	.80, .80
	.60	.63	.62	.57	.52	.49	.44	.38	.20 *(.20)*
.85	−.53, .53	−.25, .76	−.03, .88	.22, .97	.36, .99	.45, 1.00	.54, .99	.64, .97	.85, .85
	.53	.55	.53	.48	.44	.40	.36	.31	.15 *(.15)*
.90	−.44, .44	−.15, .69	.07, .83	.32, .94	.46, .98	.54, .99	.62, 1.00	.72, .99	.90, .90
	.44	.45	.43	.38	.34	.31	.28	.23	.10 *(.10)*
.95	−.31. .31	−.01, .58	.21, .75	.44, .89	.57, .95	.64, .97	.72, .99	.81, 1.00	.95, .95
	.31	.31	.29	.26	.23	.21	.18	.14	.05 *(.05)*
1.00	.00, .00	.30, .30	.50, .50	.70, .70	.80, .80	.85, .85	.90, .90	.95, .95	1.00, 1.00
	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note: r_{ay} is the correlation between the desired measure (*a*) and the dependent variable (*y*); r_{ab} is the convergent validity between the desired measure and the proxy (*b*). The body of the table lists feasible values for r_{by}, the correlation between the proxy and the dependent variable. For each combination of r_{ay} and r_{ab}, the first row in each cell lists the minimum, maximum possible values for r_{by}. The second row of each cell lists the maximum discrepancy between r_{ay} and r_{by}, and when it is relevant the minimum discrepancy is shown in italics in parentheses.

Hadavas, Hoffman, and Pretz (2007) report that given the value of two correlations (r_{12}, r_{13}) in a 3 × 3 correlation matrix, the value of the third correlation (r_{23}) can possess a range of possible values bounded by their Equation 1 (p. 55):

$$r_{12}r_{13} - \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)} \leq r_{23} \leq r_{12}r_{13} + \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}. \tag{1}$$

For our purposes, Measure 1 = X_a, 2 = X_b, and 3 = Y, so r_{12} assumes the value of a known convergent validity r_{ab}, r_{13} is a known value of r_{ay}, and r_{23} represents the potential value of the unknown relationship of the proxy with outcome Y (r_{by}). By substituting elements, the formula above can then be used to estimate the range of possible values of r_{by} for given values of r_{ab} and r_{ay}, as follows:

$$r_{ab}r_{ay} - \sqrt{(1 - r_{ab}^2)(1 - r_{ay}^2)} \leq r_{by} \leq r_{ab}r_{ay} + \sqrt{(1 - r_{ab}^2)(1 - r_{ay}^2)}. \tag{2}$$

Table 1 reports the results of an analysis of the potential range of values of r_{by} given varying levels of r_{ab} and r_{ay}. For each combination of r_{ab} and r_{ay}, three values are reported: the maximum mathematically possible value of r_{by}, the minimum possible value, and the maximum possible deviation from r_{ay}. As shown in the table, when $r_{ab} = 1$, the values of r_{by} equal those of r_{ay}. Conversely, when $r_{ab} = 0$, the value of r_{by} is not restricted—it can assume any value from $r = –1.0$ to $r = +1.0$. Between these extremes, the range of possible values of r_{by} becomes more restricted as the absolute values of r_{ab} and r_{ay} increase. However, it is interesting to note that even at what would be considered conventionally very high levels of convergent validity (i.e., $r = .90$), the range of possible values for r_{by} can differ substantially from r_{ay}. For example, for $r_{ab} = .90$ and $r_{ay} = .50$, the possible values of r_{by} range from $r = .07$ to $r = .83$, a range of .76 correlation units with a maximum possible difference from r_{ay} of .43. Even when $r_{ab} = .95$, the maximum possible difference between r_{ay} and r_{by} is $r = .29$.

These analyses clearly indicate that dramatic differences in research findings are possible even at levels of convergent validity that current research convention would consider high. These values, though, represent the upper and lower bounds of potential differences in findings. It is not clear that these values are ever reached in real data. In the current study, we develop data that offer insight into the variation in research findings that can be expected in actual research. Drawing from published research findings, we examine the impact of differing levels of convergent validity on the magnitude of research results.

Method

The purpose of this study is to examine how the degree of convergence between alternative measures affects research results and substantive conclusions. We examined the extent that weaker convergent validity as represented by weaker correlations between two measures X_a and X_b (i.e., r_{ab}) results in increasing divergence in the magnitudes of correlations between these two measures and a third measure Y (i.e., r_{ay} r_{by}). To examine these effects, we conducted a review of zero-order correlation matrices in published research to find instances where any two measures (*A, B*) were correlated at 1 of 10 targeted levels of association. Once these r_{ab} correlations were identified, we gathered data on all correlations between these two measures, X_a and X_b, with every other measure (Y_i) reported in the study's zero-order correlation matrix. We sought to identify at least 300 pairs of correlations (i.e., r_{ayi}, r_{byi}) for each targeted level of r_{ab}.

In our search for data, we conducted a manual review of published articles from the period 2006 to 2008 in several leading journals. These journals included *Academy of Management Journal, Journal of Applied Psychology, Journal of Management, Personnel Psychology,* and *Strategic Management Journal.* Journals were selected to provide a broad sampling of the measures found in the management literature and to ensure the quality and accuracy of reported data. We began by reviewing the most recent issues of each journal available, (a) searching for articles that included correlation matrices and (b) searching within correlation matrices for correlations of the magnitudes identified for our analysis.

We examined 10 levels of association for r_{ab} representing the full range of correlations between perfect association and perfect independence. Because decisions about substitutes are more likely to involve higher levels of convergence, we narrowed the distance between increments as the values approached $r = 1.0$. The levels we examined were correlations of .10, .30, .50, .60, .70, .75, .80, .85, .90, and .95. Given the difficulty in finding sufficient numbers of data examples for each specific level of r_{ab}, we included data for correlations that were +/–.01 from the target correlation. For example, $r = .10$ data included r_{ab} relationships ranging from .09 to .11. To simplify the analysis, we incorporated data only for positive r_{ab} correlations. Our target was to identify at least 300 pairs of correlations for each level of convergent validity. We ceased actively searching for a given target correlation when at least 300 pairs were identified. This was possible in all circumstances, except $r_{ab} = .95$ ($N = 237$). The largest correlations were the most difficult to find and resulted in the most extensive searches.

We chose to examine observed correlations rather than correlations corrected for measurement error. We recognize that error of measurement alone may account for attenuation of observed correlations. Doing so would attempt to separate the effects of error of measurement on convergent validity from the effects of construct validity. Our decision was based on three factors. First, in many instances the source data did not report a reliability estimate. Second, in many instances it was not possible to appropriately partition attenuation due to error of measurement from attenuation due to differences in construct validity. Third, developing better proxies will ultimately require researchers to attend to both construct validity and error of measurement.

Results

In total, data for 3,684 r_{ay}, r_{by} correlation pairs were collected. Sample sizes in the studies from which our data are drawn ranged from $N = 203$ to $N = 73{,}000$, with a median sample size of $N = 740$. This is consistent with sample sizes found in other reviews of the literature.

For each of the 10 levels of r_{ab} we conducted two analyses. First we constructed a bivariate plot of magnitudes of the correlation pairs (i.e., with r_{ay} on the x-axis and r_{by} on the y-axis). For data in which these two values are identical, points in the bivariate plot rest on the line $y = x$. As the values in correlation pairs diverge, points become more distant from this line. Examples of these plots are reported in Figure 2. A perusal of these graphs demonstrates that as the magnitude of r_{ab} is closer to 1.0, the plots more closely approximate $y = x$. Second, we determined the absolute value of the difference between the value of r_{ay} and r_{by} for each pair. We then calculated the number (and percentage) of the absolute differences between correlation pairs that fell in six difference ranges (i.e., differences of r = .00-.02, .03-.05, .06-.10, .11-.20, .21-.30, and > .30). These data were then used to construct bar charts showing the relative distribution of differences between these categories.

Table 2 presents the distribution of these differences between the magnitudes of r_{ay} and r_{by} for each level of r_{ab}. As shown, as the magnitude of r_{ab} declines, the percentage of relationships with larger differences increases. For instance, for $r_{ab} = .95$, 91% of the differences are .05 or less, with only 2% of pairs producing differences greater than .10. By the time $r_{ab} = .70$, the percentage of differences that are .05 or less drops to 43%, whereas the percentage of differences in the range of .11-.20 climbs to 30%. As r_{ab} decreases further, the magnitudes of differences continue to increase. For $r_{ab} = .50$, only 31% of differences are .05 or less, 44% of differences are greater than .10, and 17% are greater than .20. For $r_{ab} = .30$, only 23% of differences are .05 or less, 55% of differences are greater than .10, and 28% are greater than .20. As noted earlier, Bedeian et al. (2009) argue that differences in correlations of $r = .10$ or greater are likely to represent practically significant differences in effect size. As Table 2 shows, for $r_{ab} < .70$, there is at least one chance in five that the absolute value of the findings using the substitute will yield results more than $r = .10$ different from those produced by the desired measure.

Discussion

The use of proxy or substitute measures is common in organizational research. Yet the consequences of less-than-perfect convergent validity on research findings have not been clearly articulated. A critical issue for advancing literatures and ensuring confidence in findings is understanding the extent that less-than-perfect convergent validity affects results. The results of the current study demonstrate that even relatively modest departures from perfect convergent validity create the possibility for substantial differences in research findings. In our analysis, even for convergent validities as strong as $r = .90$ as many as 5% of findings for proxies may differ by more than $r = .10$ from those for the desired measure. This number grows to more than 28% for convergent validities of $r = .80$. As convergent validity falls below $r = .70$, the percentage of findings that differ by more than $r = .10$ climbs to 30%, and differences as large as $r = .20$ occur 9% of the time. As convergent validities continue to fall, the magnitude and frequency of divergent findings increase and, by extension, so does the risk for inappropriate interpretations of findings.

To date, the general guidance to researchers is to select measures with stronger convergent validity. Our data confirm that wisdom, but go further. Based on these data, statements like "some convergent validity is better than none" can reasonably be challenged. These data indicate that a proxy

Figure 2. A comparison of the relative distribution of differences magnitude of the values of r_{ay} and r_{by} at selected levels of r_{ab} (.90., .70, .50, .30, and .10)

Note: Each point in the diagram is the plot of a point in coordinate space (r_{ay}, r_{by}). If there were no differences in the value of r_{ay} and r_{by} correlations, the points would fall on the line $y = x$. The accompanying histograms provide a depiction of the number of correlation pairs whose absolute differences in magnitude fall in each of indicated ranges.

Table 2. Analysis of Absolute Differences in r_{ay} and r_{by} Across Correlation Pairs at Different Levels of Convergent Validity

Convergent Validity	*N*	*M*	*SD*	Max	Distribution of Absolute Differences					
					.00-.02	.03-.05	.06-.10	.11-.20	.21-.30	>.30
.95	237	.02	.03	.26	174 (73)	43 (18)	15 (06)	4 (02)	1 (00)	0 (00)
.90	312	.04	.04	.42	122 (39)	115 (37)	63 (20)	11 (04)	0 (00)	1 (00)
.85	325	.06	.11	1.21	113 (35)	79 (24)	87 (27)	40 (12)	2 (01)	4 (01)
.80	325	.09	.08	.64	85 (25)	72 (22)	75 (23)	72 (22)	10 (03)	11 (03)
.75	431	.08	.08	.88	96 (22)	98 (23)	116 (27)	94 (22)	21 (05)	4 (01)
.70	345	.09	.09	.76	84 (24)	67 (19)	92 (71)	71 (21)	19 (06)	12 (03)
.60	384	.11	.14	1.49	67 (18)	86 (24)	76 (21)	83 (23)	37 (10)	15 (04)
.50	377	.13	.12	.95	53 (14)	67 (17)	98 (25)	103 (26)	40 (10)	27 (07)
.30	347	.16	.16	.86	40 (11)	44 (12)	78 (22)	96 (27)	49 (14)	51 (14)
.10	525	.18	.18	1.18	51 (10)	70 (14)	78 (16)	111 (23)	82 (17)	100 (20)

Note: *N* is the number of correlation pairs examined; *M* is the mean and *SD* is the standard deviation of differences in the magnitude of r_{ay} and r_{by}. *Max* is the largest value found between the two correlations. In the columns labeled *distribution of absolute values,* the number is the actual number of correlation pairs where the absolute value of the difference falls in the identified range. The number in parentheses is the percentage of the total *N* represented by the preceding number.

with convergent validity less than $r = .50$ cannot reasonably be substituted for the desired measure. At this level of convergence, nearly half of the results for proxies differ by more than $r = .10$.

Existing levels of convergent validity in many contemporary literatures are modest. As our results demonstrate, modest levels of convergent validity slow research progress. The results of studies that use weakly converging proxies will confound, rather than contribute to, our understanding of important organizational phenomena by increasing differences in findings across studies. As shown in the debate surrounding the risk propensity of entrepreneurs (Miner & Raju, 2004; Stewart & Roth, 2001, 2004), the potential for differences in the magnitudes of findings across studies due solely to difference in construct validity exists, and those differences can be substantial. This issue is especially acute in the selection of dependent variables. The relatively weak associations often found between alternative dependent variables (Combs et al., 2005) suggest that research results for these different variables are not comparable. Differences in construct validity—signaled by poor convergent validity—increase variability in the findings of meta-analytic reviews. Greater variability in effect size estimates reduces the generalizability of results and precludes the use of point estimates of the magnitude of relationships between constructs (Carlson & Ji, 2011).

Theories are challenged when they are confronted with data about a phenomenon they cannot explain. However, data that do not represent the theorized construct cannot be expected to conform to its expectations. When measures only marginally capture intended constructs, we run a high risk of convicting the innocent (theory) while praising the unworthy. In addition, the ambiguity introduced by marginally convergent proxies further obscures the consistencies and patterns upon which understanding grows. Meehl (1990) argued for developing more precise models of organization systems and stronger theory—that which could generate point estimates of expected outcomes rather than simply directional hypotheses. This approach places greater emphasis on the substantive findings of studies. Applying this more stringent standard suggests that substitute measures should possess stronger levels of convergent validity, perhaps as high as $r = .85$, to increase the chances that the magnitudes of findings between studies is substantively the same. For the correlations examined in this study, we defined that level as $\Delta r = .10$. Greater precision offers more precise data from which to build theory and allows for the severe tests of hypotheses, both of which support the development of better theory.

Both theory and research practice in an area of study evolve over time. Early in the study of a phenomenon, theory is often described as developing or emergent. At this point, theory may offer only very broad or ambiguous definitions of constructs and speak only generally to the nature of relationships. Measurement at this stage is difficult. A first step is to move toward theory that offers clearer definitions of key constructs and relationships, remembering that measures and statements of theory will likely change. Attempts to identify the best measures of the constructs lead to insights that advance theory development. A primary challenge for researchers is to recognize instances when proxies fail to demonstrate strong convergent validity. In emerging theories, these discontinuities create conceptual tension that will lead to insights and refinements of theory and construct definitions.

The analysis conducted here also applies to formative measures. Formative measurement is an alternative to the reflective measurement discussed in classical measurement theory (Edwards & Bagozzi, 2000). Formative measurement does not begin with a construct, so there is no construct definition against which to evaluate measures. Consequently, the notion of construct validity does not exist in formative measurement. Instead, the ‘‘definition’’ of the measure is defined by the inputs and combinatorial rule used to construct it. In the terminology of Figure 1, formative approaches to measurement have no χs or ψs, just Xs and Ys. An example of a formative measure is socioeconomic status, which is calculated from a predefined set of inputs combined according to a specific formula into an index. If the inputs or the rule for combining the inputs to a formative measure are changed, a new measure is formed. Utility, rather than correspondence to theory, is the primary determinant of value for formative measures. Consequently, an alternative formative measure is preferred if it predicts an outcome more effectively. A reasonable question for researchers using formative measures is whether slight modifications to the component data or formula used substantively change the measure—are data based on these different calculations substitutable? This question should be answered by determining the convergent validities among measures based on the alternative calculations. If the resulting measures are not highly correlated, the alternative calculations cannot be substituted without consequence.

In sum, based on this analysis and discussion, we offer two recommendations for research practice. First, renewed rigor is necessary to identify proxy measures with the highest levels of convergent validity possible. Toward this end, researchers are encouraged to report evidence of convergent validity when it exists and to discuss its potential implications for their findings. The effects of poor convergent (construct) validity are not easily detected in individual studies, particularly those that employ a single proxy. Thus, reporting evidence of convergent validity for proxies aids in the appropriate interpretations of results. Second, convergent validity data are critical to making sound judgments concerning proxies. Whenever possible, researchers are encouraged to include more than one alternative measure in their research designs or to undertake studies specifically designed to develop data for the convergent validity of potential proxies. These efforts are important to enhancing confidence in findings, growing precision within literatures, and developing strong theory.

Study Limitations

Our findings are based on the sample of relationships we gathered. The number of correlation pairs and the size of the samples they were drawn from give us greater confidence in our findings. Incorporating more data would offer more precision to our results but would not substantively change our conclusions. Further, it is possible that our data contain transcription errors in the reporting of the correlation matrices within the published studies. Some values in the plots for Figure 2 are sufficiently deviant to suggest potential errors. Yet, all values reported fall within the range of mathematically possible correlations as reported in Table 1. Therefore, we included these data points in our analysis. However, excluding them would not have substantively changed our results or interpretations.

Our analysis focused on the correlation coefficient as an indicator of convergence. Although this is sufficient to understand the impact of proxies in correlational analyses, it is not sensitive to differences in means between proxies. We also do not examine the effects of common method variance on our findings. When common method variance is present, it is possible that observed correlations between measures overstate true levels of convergent validity. Shared variance attributable to common methods would by definition not be related to the construct. As a result, high convergent validity would not necessarily produce highly convergent associations with third variables. We chose not to examine common method variance in the study because existing evidence suggests that low levels of convergent validities, rather than those that are too high, are currently of greater concern. To the extent that common method variance does exist at meaningful levels in the measures we examined, our results would offer conservative estimates of the magnitude and frequency of the divergence of results.

Finally, our analysis did not focus on specific constructs. This approach led to some concern that the results of these analyses would be different if we focused on evidence for specific constructs. Although doing so would offer greater insight into specific proxies, this approach is limited in other ways. First, many proxies do not produce correlations of the specific magnitudes we sought to examine in this study. Adding this restriction to the analysis conducted here would require large bodies of empirical data for specific constructs that are unlikely to exist in any published literature. Second, even if these data existed, the resulting findings would be less generalizable. In the end, the mathematical properties of convergent validity are demonstrated irrespective of the underlying constructs these data were intended to represent.

Conclusion

The results of this analysis provide new information on the risks of weak convergent validity for the interpretation, accumulation, and application of research findings. These findings offer researchers specific numerical estimates of the likelihood that results will vary and the likely extent of variation for convergent validities ranging from $r = .10$ to $r = .95$. These data should be used to guide future measure selection and interpretation of results. Further, data on convergent validity contribute to the body of evidence upon which the construct validity of our measures is judged. Therefore, we encourage researchers to include alternative proxy measures in research designs and contribute to reporting data on the convergent validity of proxy measures whenever possible.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Acquaah, M. (2007). Managerial social capital, strategic orientation and organizational performance in an emerging economy. *Strategic Management Journal*, *28*, 1235-1255.

Bakker, A. B., Demerouti, E., & Dollard, M. F. (2008). How job demands affect partners' experience of exhaustion: Integrating work–family conflict and cross-over theory. *Journal of Applied Psychology*, *93*, 901-911.

Bedeian, A. G., Sturman, M. C., & Streiner, D. L. (2009). Decimal dust, significant digits, and the search for stars. *Organizational Research Methods*, *12*, 687-694.

Bercovitz, J., & Mitchell, W. (2007). When is more better? The impact of business scale and scope on long-term business survival, while controlling for profitability. *Strategic Management Journal*, *28*, 61-79.

Bommer, W. H., Dierdorff, E. C., & Rubin, R. S. (2007). Does prevalence mitigate relevance? The moderating effect of group-level OCB on employee performance. *Academy of Management Journal*, *50*, 1481-1494.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). One the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology*, *48*, 587-605.

Bowling, N. A., Hendricks, E. A., & Wagner, S. H. (2008). Positive and negative affectivity and facet satisfaction: A meta-analysis. *Journal of Business Psychology*, *23*, 115-125.

Boyd, B. K., Gove, S., & Hitt, M. A. (2005). Consequences of measurement problems in strategic management research: The case of Ahmihud and Lev. *Strategic Management Journal*, *26*, 367-375.

Budden, M., Hadavas, P., Hoffman, L., & Pretz, C. (2007). Generating valid 4×4 correlation matrixes. *Applied Mathematics E-notes*, *7*, 53-59.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Carlson, K. D., & Ji, F. X. (2011). Citing and building on meta-analytic findings: A review and recommendations. *Organizational Research Methods*, *14*, 696-717.

Chen, W. R., & Miller, K. D. (2007). Situational and institutional determinants of firms' R&D search intensity. *Strategic Management Journal*, *28*, 369-381.

Combs, J. G., Crook, T. R., & Shook, C. (2005). The dimensionality of organizational performance and its implications for strategic management research. In D. Ketchen & D. Bergh (Eds.), *Research methodology in strategy and management* (pp. 259-286). San Diego, CA: Elsevier.

Earle, J. S., Spicer, A., & Peter, K. S. (2010). The normalization of deviant organizational practices: Wage arrears in Russia, 1991-1998. *Academy of Management Journal*, *53*, 218-237.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and relationship between constructs and measures. *Psychological Methods*, *5*, 155-174.

Fey, C. F., & Furu, P. (2008). Top management incentive compensation and knowledge sharing in multinational corporations. *Strategic Management Journal*, *29*, 1301-1324.

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment. *Journal of Personality and Social Psychology*, *61*, 457-476.

Glass, G., & Collins, J. (1970). Geometric proof of the restriction on the possible values of r_{xy} when r_{yz} and r_{yz} are fixed. *Educational and Psychological Measurement*, *30*, 37-39.

Gong, Y., Law, K. S., Chang, S., & Xin, K. R. (2009). Human resource management and firm performance: The differential role of managerial affective and continuance commitment. *Journal of Applied Psychology*, *94*, 263-275.

Guler, I., & Guillen, M. F. (2010). Home country networks and foreign expansion: Evidence from the venture capital industry. *Academy of Management Journal*, *53*, 390-410.

Guthrie, J. P. (2001). High-involvement work practices, turnover and productivity: Evidence from New Zealand. *Academy of Management Journal*, *44*, 180-190.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Jackson, D. N. (1976). *Personality inventory manual.* Goshen, NY: Research Psychologist Press.

Judge, T. A., Jackson, C. L., Shaw, J. C., Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology*, *92*, 107-127.

Kale, P., & Singh, H. (2007). Building firm capabilities through learning: The role of the alliance learning process in alliance capability and firm-level alliance success. *Strategic Management Journal*, *28*, 981-1000.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individual's fit at work: A meta-analysis of person–job, person–organization, person–group, and person–supervisor fit. *Personnel Psychology*, *58*, 281-342.

Larraza-Kintana, M., Wiseman, R. M., Gomez-Mejia, L. R., & Welbourne, T. M. (2007). Disentangling compensation and employment risks using the behavioral agency model. *Strategic Management Journal*, *28*, 1001-1019.

Lee, G. K. (2007). The significance of network resources in the race to enter emerging product markets: The convergence of telephony communications and computer networking, 1989–2001. *Strategic Management Journal*, *28*, 17-37.

Maurer, T. J., & Andrews, K. D. (2000). Traditional, Likert, and simplified measures of self-efficacy. *Educational and Psychological Measurement*, *60*, 965-973.

Maurer, T. J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology*, *83*, 324-329.

McDonald, M. L., & Westphal, J. D. (2010). A little help here? Board control, CEO identification with the corporate elite, and strategic help provided to CEOs at other firms. *Academy of Management Journal*, *53*, 343-370.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable [Monograph]. *Psychological Reports*, *66*, 195-244.

Miner, J. B., & Raju, N. S. (2004). Risk propensity differences between managers and entrepreneurs and between low and high growth entrepreneurs: A reply in a more conservative vein. *Journal of Applied Psychology*, *89*, 3-13.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873-900.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1-18.

Nunnally, J. C. (1967). *Psychometric theory.* New York, NY: McGraw-Hill.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Blume, B. D. (2009). Individual- and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology*, *94*, 122-141.

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practice, trends and implications for future research. *Academy of Management Journal*, *43*, 1248-1264.

Schwab, D. (1980). Construct validity in organizational behavior. *Research in Organizational Behavior*, *2*, 3-43.

Short, J. C., Ketchen, D. J., Jr., Shook, C. L., & Ireland, R. E. (2010). The concept of opportunity in entrepreneurship research: Past accomplishments and future challenges. *Journal of Management, 36,* 40-65.

Spangler, W. E. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin*, *112*, 140-154.

Stanley, J., & Wang, M. (1969). Restrictions on the possible values of r_{12}, given r_{13} and r_{23}. *Educational and Psychological Measurement*, *29*, 579-581.

Stewart, W. H., & Roth, P. L. (2001). Risk propensity differences between entrepreneurs and managers: A meta-analytic review. *Journal of Applied Psychology*, *86*, 145-153.

Stewart, W. H., & Roth, P. L. (2004). Data quality affects meta-analytic conclusions: A response to Miner and Raju (2004) concerning entrepreneurial risk propensity. *Journal of Applied Psychology*, *89*, 14-21.

Toh, S. M., Morgeson, F. P., & Campion, M. A. (2008). Human resource configurations: Investigating fit with the organizational context. *Journal of Applied Psychology*, *93*, 864-882.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* (Vol. 3). Thousand Oaks, CA: Sage.

Wall, T. D., Michie, J., Patterson, M., Wood, S. J., Sheehan, M., Cleeg, C. W., & West, M. (2004). On the validity of subjective measures of company performance. *Personnel Psychology*, *57*, 95-118.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, *76*, 820-838.

Bios

Kevin D. Carlson is an Associate Professor of Management in the Pamplin College of Business at Virginia Tech. He has published research on a wide variety of topics related to the measurement and evaluation of individual, process and organizational effectiveness. His current research interests center on how research methodology contributes to research progress.

Andrew O. Herdman is an Assistant Professor of Management at East Carolina University. His research and teaching are focused in the areas of strategic human resource management, negotiation and organizational behavior.