Communications of the Association for Information Systems

Volume 47 Article 64

12-31-2020

Empirical Test Guidelines for Content Validity: Wash, Rinse, and Repeat until Clean

Kurt Schmitz
Georgia State University, kschmitz1@gsu.edu

Veda C. Storey
vstorey@gsu.edu

Recommended Citation
Schmitz, K., & Storey, V. C. (2020). Empirical Test Guidelines for Content Validity: Wash, Rinse, and Repeat until Clean. Communications of the Association for Information Systems, 47, pp-pp. https://doi.org/10.17705/1CAIS.04736

Communications of the Association for Information Systems

Research Article DOI: 10.17705/1CAIS.04736 ISSN: 1529-3181

Empirical Test Guidelines for Content Validity: Wash, Rinse, and Repeat until Clean

Kurt Schmitz
J. Mack Robinson College of Business
Georgia State University
kschmitz1@gsu.edu

Veda C. Storey
J. Mack Robinson College of Business
Georgia State University
vstorey@gsu.edu

Abstract:

Empirical research in information systems relies heavily on developing and validating survey instruments. However, researchers' efforts to evaluate content validity of survey scales are often inconsistent, incomplete, or unreported. Thjs paper defines and describes the most significant facets of content validity and illustrates the mechanisms through which multi-item psychometric scales capture a latent construct's content. We discuss competing methods and propose new methods to assemble a comprehensive set of metrics and methods to evaluate content validity. The resulting recommendations for researchers evaluating content validity emphasize an iterative pre-study process (wash, rinse, and repeat until clean) to objectively establish "fit for purpose" when developing and adapting survey scales. A sample pre-study demonstrates suitable methods for creating confidence that scales reliably capture the theoretical essence of latent constructs. We demonstrate the efficacy of these methods using a randomized field experiment.

Keywords: Content Validity, Latent Construct, Clarity, Relevance, Internal Congruence, External Congruence, Contamination, Adequacy, Indicator Sufficiency, Indicator Parsimony, Dependability, Reliability, Stability, Psychometric Scale.

This manuscript underwent peer review. It was received 06/25/2019 and was with the authors for ten months for three revisions. Alan R. Dennis served as Associate Editor.

1 Introduction

Research in the information systems (IS) field involves imagining, describing, and testing theories related to multiple phenomena. Theorized phenomena are idealized concepts (something the researcher has imagined) that should exist in the real world. They have specific definitions that establish boundaries of content domains. Although researchers can directly measure some concepts they use to understand our world (e.g., how frequently a user makes a phone call or how much time elapses between two events), many noumena[1] involve attitudes, perceptions, emotions, and other cognitive ideas that one cannot observe. Theories in the behavioral sciences broadly, and the IS field specifically, abound with *latent constructs,* which researchers use when they cannot directly measure a theorized noumenon's properties. The behavioral sciences have embraced the process of modeling noumenon indirectly by measuring observable indicators. Researchers measure properties of noumenon using multi-item psychometric surveys. Collectively, data from multiple items establish a measure of an inferred latent construct.

A fundamental tenant of empirical research is validity. Consequently, scientific inquiry attempts to address many aspects of validity (Trochim, 2006). External validity deals with generalization: can the conclusions based on one set of observations generalize to other persons, places, and times? Conclusion validity deals with documenting a non-random relationship between two constructs: is the correlation statistically significant or is it plausibly a random coincidence? Internal validity deals with causality: does one noumenon or event cause a second noumena or event?

Construct validity and content validity become relevant when researchers measure a latent construct indirectly using manifest indicators. Construct validity embraces various issues using statistical properties of measurement scores. Researchers assess it after collecting primary data. The IS literature has embraced many statistical procedures and heuristic criteria for construct validity (MacKenzie, Podsakoff, & Podsakoff, 2011; Straub, Boudreau, & Gefen, 2004; Urbach & Ahlemann, 2010). Content validity deals with the question: are we measuring the content (the noumena) we intend to measure? In the IS field, "the most commonly employed evaluation of this validity is judgmental and is highly subjective" (Straub et al., 2004, p. 387). However, science abhors subjective judgments and prefers stable and replicable evaluation methods. While researchers routinely examine aspects of construct validity using empirical and analytical methods (Boudreau, Gefen, & Straub, 2001; Ringle, Sarstedt, & Straub, 2012), they often neglect content validity (MacKenzie et al., 2011; Hoehle & Venkatesh, 2015).

Establishing many aspects of validity is predicated in decisions made during the study's design. A best practice for studies attempting to establish causality (conclusion validity) involves planning data collection and treatments in a temporally controlled sequence. A different concern that applies to cross-sectional study designs concerns the need to mitigate common method bias. Investigators choosing to implement the Marker technique (Williams, Hartman, & Cavazotte, 2010) must develop and include a marker variable and associated items in the survey design. In both situations, researchers establish the foundation when designing a study.

Achieving content validity also requires researchers to carefully plan and craft measurement instruments during study design. Content validity processes involve iteratively assessing and revising survey instruments. The techniques involve feedback-collection exercises during construct development. These exercises often involve much smaller participant pools than exercises to gather data for a main study. We refer to primary study scores as **study data** and to the feedback that one collects during instrument development as **panel data**. Furthermore, we refer to participants in a primary study as informants and participants in an instrument development pre-study as jurors. Researchers sometimes assemble informants for a pilot study, which they conduct after the pre-study but before the primary study. Pilot studies allow researchers to evaluate the measurement model for construct validity with a smaller informant sample prior to assembling a larger cohort to test conclusion validity. Instrument development exercises are iterative and may involve multiple different panels just as some study designs involve multiple data-collection events (e.g., longitudinal designs) or multiple subject pools (e.g., triangulation designs).

[1] Much of the relevant literature uses the more common term "phenomenon". Phenomena refer to directly observable facts or events. In this paper, we focus on unraveling the intricacies of latent constructs. Therefore, we adopt the less common but more precise term "noumenon" as the label for an object or event that exists beyond one's ability to perceive via the senses.

Construct validity provides various quantitative metrics (e.g., internal consistency reliability) that researchers sometimes use to inform aspects of content validity. However, assessing content validity with study data constitutes a problem for several reasons:

- Content validity tests expose issues that researchers should address prior to collecting data for conclusion validity to remove unnecessary measurement error that make conclusion tests unstable, unreliable, or misleading.
- Achieving content validity is an iterative process: researchers should fix each problem that arises and revalidate the updated instrument (or, in other words, wash, rise, and repeat until clean).
- Researchers may need a large minimum sample size to achieve the power necessary to test many hypotheses. Using study data to inform content validity presents tremendous risk because researchers must discard the data when they discover validity flaws late in the study. For many studies, it is prohibitively expensive to assemble multiple large groups of subjects and repeat the study.
- Some studies (e.g., longitudinal studies) require researchers to collect data at specific times that align with the study protocol. In most field settings, researchers cannot simply stop, adjust the survey scales, and repeat a data-collection exercise.

Primary data-collection exercises involve pools of informants who evaluate a noumenon of interest. A juror evaluating the measurement instrument focuses on their interpretation of item(s), not their personal relationship with the noumenon. The object of investigation is different, so the primary data-collection exercise will have poor content validity from which to judge content validity. The primary data-collection exercise inherently contains significant measurement error about the measurement instrument's quality[2]. We examined recent IS literature and found that many researchers employ inconsistent and incomplete methods to address content validity.

The objectives of this paper are to: 1) identify and define important facets of content validity 2) summarize the current state of content validity assessment in the IS literature, 3) identify specific empirical methods for testing each facet of content validity, and 4) demonstrate (using an example) the core processes and their efficacy. This paper contributes to the literature by refining the multi-step development process for survey instruments to better incorporate content validity and, thus, provide scholars who seek to evaluate content validity a comprehensive set of tools for doing so.

This paper proceeds as follows. Section 2 presents the main concepts underlying content validity and reviews the state of content validity in the information systems literature. Section 3 defines each facet of content validity, along with prescriptive methods, and test metrics. Section 4 presents quantitative metrics and summarizes their appearances in the content validity literature. Section 5 provides an example of content validity rating mechanics and qualitative metrics calculations. This section also details a field experiment that demonstrates the efficacy of these methods. Section 6 discusses the study's implications and potential topics for future research. Section 7 concludes the paper. Supplemental information on testing, related concepts, and survey instruments is provided in the appendices.

2 Content Validity Concepts

Convergent validity, discriminant validity, and reliability describe how well items represent "something" that one cannot directly observe, but do not indicate whether the "something" is the theorized noumenon. Content validity deals with the degree to which items capture the theoretical essence of what they propose to measure. When the operational definition accurately portrays the theorized definition (i.e., the inferred latent construct closely represents the real noumenon), then the measures have content validity. When the inferred latent construct diverges from the real construct, the study data contains measurement error. "By maximizing content validity, the predictive validity of a test is enhanced" (Sireci, 1998, p. 107).

Improving content validity is an exercise to reduce measurement error. The processes suitable to assess and improve validity depend on the methods researchers used to measure noumena. This paper focuses

[2] An example of this involves *External Congruence* and *Adequacy* that will be introduced in the next section. Data from the primary study lacks Adequacy for evaluating External Congruence as its scope is limited to constructs in the theoretical model. The exercise is not capable of assessing the constellation of orbiting constructs that may contaminate an item.

on capturing data from survey instruments (multi-item psychometric scales in particular). The mechanics and nature of psychometric scales provide a necessary foundation for establishing and assessing their content validity.

2.1 Psychometric Scales

Science has its own jargon and ways to define noumena with precision. Study participants with a layman's grasp of language would find a theoretical construct using scientific jargon abstruse. A lay study informant could be confused by the jargon and answer based on an incorrect interpretation. Rather than translating a theoretical construct's precise definition from scientific jargon to common language as a single question (an impossibly complex task), researchers devise multiple questions to measure an implied construct. They present each question in simple language to ask about aspects or facets of the focal construct. Although no single question alone describes the construct, the questions collectively allow researchers to establish a proxy measure of the theorized construct.

Researchers must design questionnaires with care because they can influence informants' interpretation via 1) the instructions they provide, 2) the wording of questions/statement (often called indicators or items), 3) the type of answer choices (continuous values such as age in whole numbers, categorical values such as colors blue/green/red, or ordinal values such as low/medium/high), and 4) the labels applied to answer choices (called anchors for ordinal choices). Researchers commonly use items that provide a statement and then collect ordinal data with five to seven rating options using anchors from strongly disagree to strongly agree (see Figure 1). This "Likert"-style question offers rating choices along a bipolar continuum with a neutral position in the middle. Researchers can also use other types of questions such as continuous variable, dichotomous (true/false), nominal (unordered choices), semantic differential, and cumulative Guttman scales (Trochim, 2006). This paper does not address these other forms, although content validity considerations apply to them as well.

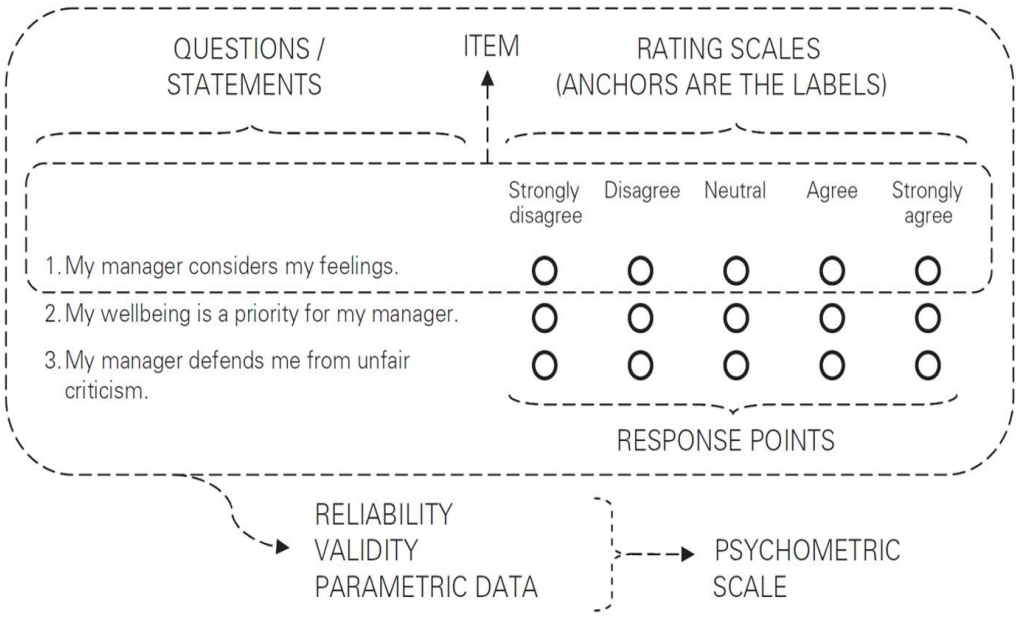

Figure 1. Psychometric Scale (Robinson, 2018)

The set of items that inform a single latent construct constitutes a multi-item psychometric scale or, more simply, scale. Most studies involve multiple constructs, each with its own scale. The collection of multiple survey scales, along with the instructions, is the survey instrument. Researchers developing and validating survey scales need to distinguish between reflective and formative latent constructs.

2.1.1 Reflective Constructs

Measuring a noumenon with reflective items involves obtaining data from a set of items that collectively characterize it. Each item asks about an observable property affected by a common underlying concept (the latent noumenon). Reflective items measure manifest properties that result from the noumenon such

that meaning emanates from the noumenon to the items (Ellwart & Konradt, 2011; Petter, Straub, & Rai, 2007). Because reflective items reveal effects from the common latent construct, the measures highly correlate with one another. Removing one item from a set of four or more may be acceptable with only a modest reduction in accuracy due to information redundancy captured across all items.

A metaphor illustrates content validity mechanisms between reflective items and the latent construct: imagine the noumena as an invisible star that one cannot directly measure. Each planet surrounding the star has properties that one can measure. The star shapes and determines these properties. Measuring properties of planets make it possible to infer information about the invisible star (such as its size or location). Identifying and measuring properties from multiple planets (multiple items) increases the dependability of the inferences that one can make about the invisible star (the latent construct). Some planets are farther from the star and, therefore, provide less or weaker information about the star. If all observable planets are a great distance from the star, then one may require information from many planets to obtain reliable inferences about the target star. Measuring a few close planets can provide information that represents the star effectively and would make less representative information from a more distant planet redundant and potentially expendable in the context of a specific study. When researchers reword an item, then they metaphorically capture information from a different planet (hopefully one that can provide more representative information). Researchers must have enough items to effectively infer the latent construct's properties but need not include so many as to fatigue the study participant and, thereby, introduce measurement error of a different sort. Building confidence in content validity requires **relevant** items; researchers should ensure that each item is highly relevant to its theorized latent construct.

High-quality questions play a role in accurately gathering data. For example, scholars suggest avoiding double-barreled questions (i.e., compound questions with an "and" conjunction) in surveys (Krosnick & Presser, 2018). In our metaphor, such questions equate to measuring two planets at the same time. Although doing so might work well when their orbits align, double-barreled questions generate chaotic information as the orbits diverge. Such questions introduce significant measurement error because researchers cannot distinguish which specific planet an informant is focused.

Scholars also suggest using simple familiar words and avoiding negations. Negations are cognitively complicated and, thus, can introduce measurement error. In our planet metaphor, negations equate to observing a planet with a dirty or foggy optical lens. As a result, informants may have only a vague idea of the target that researchers intend. Building confidence in content validity requires **clear** items; researchers should ensure that each item is precise and easy to understand.

An important aspect of content validity concerns the need to assure that study participants provide information on the intended construct. Poorly worded instructions and questions might lead an informant to provide an answer that refers to a different concept (as if the item were a planet orbiting a different, perhaps nearby, star). Scholars refer to similar, but different, latent constructs as "orbiting constructs" (Colquitt, Sabey, Rodell, & Hill, 2019, p. 1243). Following this metaphor, seemingly similar latent constructs (stars) may be in near proximity as they orbit a galaxy of noumena related to the human condition. When participants interpret an item as if it belonged to a nearby construct, they produce data that reflects the neighboring construct rather than the intended construct. Building confidence in content validity requires item-construct **congruence**; researchers should ensure each item corresponds more to the theorized latent construct than to nearby, but distinct, latent constructs.

2.1.2 Formative Constructs

To measure a latent construct with formative measures, researchers need to collect a set of indicators that either cause, or collectively produce, the latent phenomenon. The indicators jointly influence the latent construct with meaning flowing from the indicator to the construct (Ellwart & Konradt, 2011; Petter et al., 2007). Researchers often refer to formative indicators as causal indicators, which comprise a latent phenomenon's facets. Researchers collectively refer to casual indicators as an index to distinguish them from reflective scales (Diamantopoulos & Siguaw, 2006). Formative constructs derive their full meaning from their measures and, therefore, an index is not robust to omission of important indicators.

For example, a single measure of system features may reveal an IS project's success in some settings (particularly those influenced by agile development methods): how fully does the system provide the features that users desire/need? In other settings, particularly in organizations with limited resources, a second dimension, cost, is also relevant: did the project cost exceed what the organization could afford?

In this situation, a questionnaire could measure properties of features in one question and properties of cost in a second. In yet another setting, one could add the concept of time: will the security system be ready when the Olympic Games begin? Although the first two indicators (features and cost) may be fully successful, one would find it difficult to consider the project a success if it was not ready when the Olympic Games began. Formative latent constructs comprise all important indicators. A formative construct's facets are independent with little or no redundant information. Thus, omitting any one facet may incorrectly characterize the noumenon and fully invalidate the measure.

Another metaphor illustrates the challenge that researchers face in achieving content validity for formative constructs: imagine the latent construct as a hub in a wheel and the formative indicators the spokes. The theorized construct represents the real hub that one cannot see directly. Measuring the spokes allows researchers to learn the hub's properties, such as its location in the wheel. The spokes (the formative indicators) collectively reveal an inferred hub. The distance between the real hub and the inferred hub represents measurement error. Good content validity occurs in situations with a small distance between the inferred hub and the real hub. Now, extend the metaphor and imagine the spokes have an elastic property (e.g., a rubber band or bungee cord). Each spoke pulls the inferred hub in its direction. A single item (such as the system features in the example) pulls the inferred hub location (e.g., IS project success) away from the real hub location toward the rim of the wheel. When researchers use a single indicator, the inferred hub becomes indistinguishable from the indicator. While one may appropriately collect data using a single item if one can directly observe the theoretical construct, it no longer constitutes a formative latent construct. Adding a second and third indicator (such as cost and time) adds additional elastic spokes that also pull on the inferred hub. The tension from a set of well-chosen indicators will position the inferred hub with the same properties (e.g., location) as the real hub. Follow the IS project success analogy to add a fourth facet, quality. Many organizations consider system quality as equal in importance as the other facets for IS project success. Quality is orthogonal to cost, time, and features. In the analogy, reimagine the wheel on a flat plane as a sphere with a core and elastic spokes. The new indicator, being orthogonal to the others, does not lie on a flat x-y axis; rather, it lies along a third dimension, z. Adding the quality facet pulls the inferred core up (or down) the z-axis. Omitting the quality measure results in an incorrect inferred core location and an erroneous representation of the real core. An incomplete set of measures (e.g., measuring features and cost but omitting time and quality) introduces significant measurement error as the included items pull the inferred core toward themselves and away from the real location. Building confidence in content validity requires **indicator sufficiency**; researchers should ensure that all important facets of a latent construct are measured.

Researchers need to select facets to properly capture the noumenon's essence. Researchers should not simply add many indicators (spokes); rather, the indicators should proportionally represent each important facet of the latent construct. If two indicators come from closely related domains (e.g., two indicators from the cost domain with one item concerning actual cost and another cost variance), the effect equates to adding two elastic spokes from the same region of the sphere. Additional indicators pull the inferred core toward the overrepresented domain. Ideally, each indicator should represent orthogonal facets of the latent construct. To limit measurement error, researchers should include all important facets while simultaneously not overrepresenting any facet. Building confidence in content validity requires **indicator parsimony**; researchers should ensure that no facet of a formative construct is over-represented.

Clarity is as important for formative indicators as it is for reflective items. In the metaphor for formative constructs, unclear indicators contaminate the spokes' elasticity. A leading question may bias responses and give a spoke overly strong tension. An unduly complex question that confuses informants may attenuate responses and result in a spoke with relatively weak tension. Either situation allows the inferred core to float toward spokes with greater tension and away from the real core. Building confidence in content validity requires **clear** formative indicators; researchers should ensure that indicators are precise and unbiased.

2.1.3 Composite Indicator Constructs

Both reflective and formative indicators involve latent constructs. The "composite indicator" model constitutes the third type of construct. Bollen (2011, p. 360) describes this model as a variant of the formative construct but whose indicators lack causality and do not tap the same concept. Rather, they constitute a convenient collection of several variables that share a similar theme. Bollen (2011) provides a study-specific demographic composite as an example. In one study, it might involve age, gender, and race. In another study, it might involve education, income and industry employed. These contrived

constructs have no conceptual unity and no corresponding theoretical concept, so validity loses its meaning. Content validity for composites is not addressed in this paper.

2.2 Content Validity in IS Research

To reveal the state of content validity in IS research, we examined studies that employed survey techniques and were published in the last two years by leading IS journals. We identified papers in the top three IS journals recognized by *Financial Times* (*MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*). Among the remaining "basket" of eight top IS journals, we included the *Journal of the Association of Information Systems* (*JAIS*). Among these four journals, *JAIS* has the most space latitude (as measured by word and page count) and, therefore, provides the best opportunity for researchers to include content validity details should they choose to do so. In total, we examined 79 papers. Table 1 summarizes their coverage of construct validity and content validity with additional detail available in Appendix A.

Table 1. Validity Testing in IS Literature (2018-2019)

Validity assessment	Discussion	Analytical test	Suitable analytic methods
Construct validity	**77%**	**77%**	
Reflective indicator reliability	68%	68%	Indicator loading
Internal consistency	73%	73%	alpha (tau-equivalent reliability), rho/omega (composite/ congeneric reliability)
Convergent validity	61%	61%	AVE (average variance extracted)
Discriminant validity	65%	65%	CFA (chi-square difference test), cross loading, AVE > highest squared correlation with other latent variables
Formative indicator weight‡	14%	14%	Significant correlation
Collinearity‡	33%	33%	VIF
Nomological validity	33%	33%	CFA (fit measures SRMR, RMSEA, CLI, TLI, NNFI), R^2 (coefficient of determination), Q^2 (predictive relevance)
Common method bias	35%	35%	Harmon, marker variables
Content validity	**44%**	**19%**	
Construct clarity†	0%	0%	Construct rating r_{WG}
Item clarity	13%	0%	Item rating CVI, r_{WG}
Relevance (internal congruence)	18%	10%	Item matching (card sort) p_{sa}, FVI, kappa, item rating r_{WG} CVI, kappa*, htc
Contamination (external congruence)	1%	0%	Item matching (card sort), item rating c_{sv}, htd, htd*
Indicator sufficiency	0%	0%	Nominal group technique
Indicator parsimony	0%	0%	CVR, htd*
Content consistency	15%	14%	alpha (tau-equivalent reliability), rho/omega (composite/ congeneric reliability), $r^*_{WG(J)}$
Content stability	4%	0%	ICC(K), ICC(A,K)

‡ Researchers use formative constructs less frequently than reflective constructs, which explains the relatively low frequency for reporting weights and *VIF*.
† Some studies that reported efforts to resolve clarity emphasized the wording of items. Some authors report pre-study investigation included instructions, which suggests the pre-study panels may have considered construct clarity indirectly. However, no studies described any focused effort to assess construct clarity.

Over three quarters of the studies that used surveys to collect data determined construct validity to be important enough to discuss. All studies that recognized construct validity documented at least one objectively quantifiable analytic test; most documented evidence using multiple methods. Although fewer studies chose to recognize nomological validity or common method bias, studies that did so consistently documented an objectively quantifiable analytic test. In stark contrast, 44 percent of studies presented content validity as an area worth consideration. Most studies that recognized content validity reported a

pre-test effort (often involving informants not from the study team) but with unspecified feedback and analysis methods. Less than one in five studies provided objectively quantifiable analytic test for even one content validity dimension. No study reported analytical details for all aspects of content validity. This summary suggests that, although many IS scholars view content validity worthy of consideration, they perhaps lack awareness of suitable methods and tests.

3 Defining Facets of Content Validity and Quantitative Tests

Researchers often conflate content validity with face validity based on intuitive judgments rather than an explicit procedure (Johnston et al., 2014). Too often researchers assume the full burden of judging content validity, which contradicts the premise that validity is a quantitatively based judgement (Haynes, Richard, & Kubany, 1995). Evaluating content validity starts with identifying the conditions necessary for content validity. Guion (1977) provides a decomposition of content validity revealing four distinct facets. Each facet constitutes a causal indicator of content validity (Johnston et al., 2014; Haynes et al., 1995).

We present each facet below along with insights from other scholars to further clarify the challenge each represents. Researchers often conflate aspects of content validity and, thereby, mask or disregard important qualifying conditions. Figure 2 and Table 2 present precise labels for each facet.

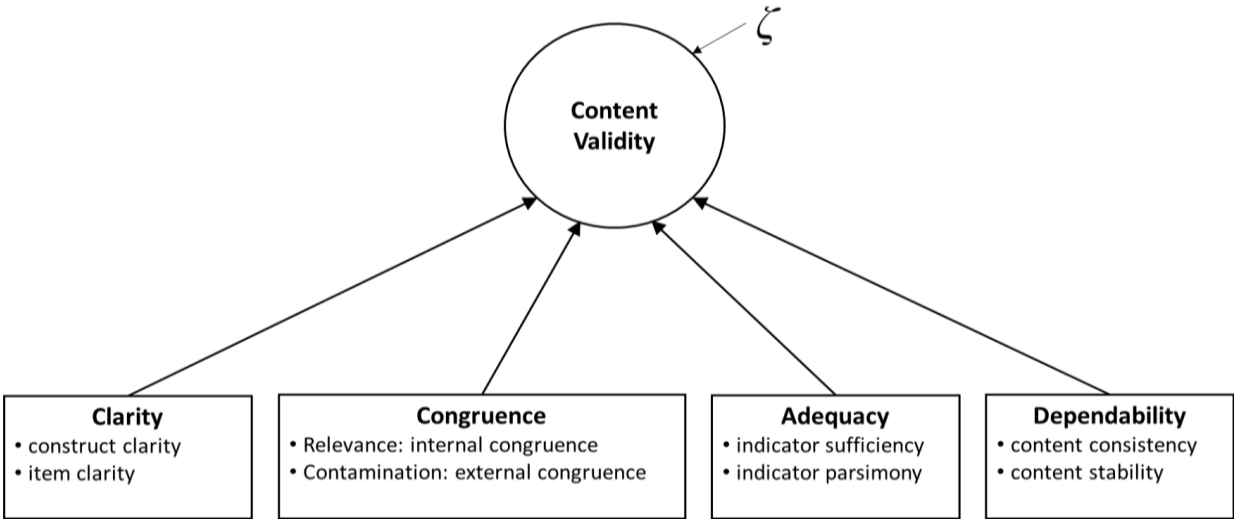

Figure 2. Facets of Content Validity

Table 2. Facets of Content Validity

Content validity		Inferences made from a scale to a noumenon are valid only to the degree to which a scale captures the theoretical essence of the noumena it is intended to measure.
Clarity	Construct clarity	A noumenon should have a generally accepted definition as it exists in a proposed nomological network.
	Item clarity	A noumenon should have an understandable operational definition as it exists in scale items.
Congruence	Internal congruence (relevance)	Items should be highly relevant to their intended noumena.
	External congruence (contamination)	Items should lack contamination from other similar noumena.
Adequacy	Indicator sufficiency	All significant facets of a noumenon should be included.
	Indicator parsimony	Each facet of a noumenon should be proportionally represented.
Dependability	Content consistency	A scale as a collective should consistently present a noumenon's content to jurors drawn from the target population.
	Content stability	A scale as a collective should have a stable interpretation by Jurors drawn from the target population.

3.1 Clarity

Clarity provides a foundation for content validity's other facets. According to Guion (1977, p. 6), "the content domain must be rooted in behavior with a generally accepted meaning". The content domain refers to a precise and specific definition of a theoretical noumenon. As Bagozzi and Phillips (1982, p. 465) state: "theoretical concepts usually consist of descriptions of phenomena provided by sentences reflecting the conceptual vocabulary of the theory". Therefore, to establish content validity, researchers first need to precisely define each latent construct as it exists in the nomological network of the proposed theory. This definition must possess generally accepted meaning, and individuals who attempt to understand the proposed theoretical model should be able to easily grasp the definition. This clear understanding must be projected into the wording of survey questions that constitute the operational definition of the latent construct. Furthermore, clarity must ultimately find its way to the study informants who see, evaluate, and respond to survey items. This clarity should exist even when these informants do not understand, or are not aware of, the proposed theory, and its nomological network. Furthermore, they should find the items clear even when they lack the necessary background to appreciate the precision with which researchers have defined a theoretical noumenon or the jargon in such a definition. In summary, clarity requires:

1) A generally accepted definition of the noumenon as it exists in a proposed nomological network (construct clarity), and
2) An understandable operational definition of the noumenon as it exists in scale items (item clarity).

Most researchers do not present their noumena's formal definitions, or a theoretical model's nomological network, to study informants. Therefore, researchers should first conduct a pre-study when developing their instrument to assess construct clarity. The pre-study should evaluate the wording of both the latent construct definition (first requirement) and the survey items in a combined form with instructions and anchors (second requirement). Gehlbach and Brinkworth (2011) propose that jurors assess item clarity (second requirement) by rating items for comprehension and understanding. Rubio et al. (2003) propose that jurors rate item clarity using these anchors: 1) item is not clear, 2) item needs major revisions to be clear, 3) item needs minor revisions to be clear, and 4) item is clear. Researchers should provide a space for comments and suggestions.

Researchers can objectively test agreement on item clarity by calculating the statistic r_{WG}. A score of 0.70 or above indicates good agreement (Brown & Hauenstein, 2005). Researchers should examine poor items for rewording and/or removal. Response values of "not clear", "needs major revision", and "needs minor revision", guide researchers regarding what action they should consider to remedy clarity problems. Researchers may also interview jurors to elicit additional insight to guide rewording.

Researchers can also use Rubio et al.'s (2003) clarity scale to assess latent construct definition. The formal latent construct definition is presented to jurors for rating using these anchors: 1) the definition is not clear; 2) the definition needs major revisions to be clear; 3) the definition needs minor revisions to be clear; and 4) the definition is clear. Again, they should provide a space for comments and suggestions. As with item clarity, researchers can objectively test agreement on construct definition clarity by calculating the statistic r_{WG}. A score of 0.70 or above supports good agreement (Brown & Hauenstein, 2005). Poor agreement warrants researchers reword the latent construct definition, while considering comments from jurors. Researchers may also interview jurors to elicit additional insight to guide rewording.

3.2 Congruence

According to Guion (1977, p. 6), "the content domain must be relevant to the purpose of measurement". Published studies discuss this content validity facet more than any other. Measures of observed items should match a latent construct's properties—a crucial requirement for reflective constructs where meaning is projected from the latent phenomenon to manifest items. Highly relevant items correspond closely with the latent construct, whereas weakly relevant items correspond only loosely.

Researchers have applied alternate labels to this concept in different settings. Hambleton (1984, p. 207) use the term item-objective congruence to capture the extent to which an item "reflects, in terms of their content, the domains from which they were derived". Anderson and Gerbing (1991, p. 732) describe substantive validity as the "extent to which that measure is judged to be reflective of, or theoretically linked to, some construct of interest". Haynes et al. (1995) use the label "relevance" to refer to elements'

appropriateness for the targeted construct and assessment function. They further note that relevance decreases with respect to the degree that items exist outside the target domain. Rubio et al. (2003) contend that an item's ability to represent the content domain as described by the theoretical definition demonstrates "representativeness". Robinson (2018, p.742) describe "conceptual fit" as the extent to which the scale matches the variable that the researcher wishes to measure. Colquitt et al. (2019, p. 1) define "definitional correspondence" as the degree to which a scale's items correspond to the construct's definition. We adopt the label "internal congruence" and the synonym "relevance" to unify the various labels in prior literature.

In addition to being relevant to their intended construct, items must be unambiguous. Latent construct definitions establish boundaries that separate one noumenon from others. According to Guion (1977, p. 6):

> *The boundaries of a domain should be clear enough that different people understanding the measurement problem at hand should be able to recognize reasonably well whether a particular item…is inside or outside those boundaries.*

Boundaries ensure that influence from other constructs do not contaminate the measure (Johnston et al., 2014). Although internal congruence indicates the extent to which an item reflects its intended construct, it does not reveal the extent to which an item might also capture other, unintended constructs (Anderson & Gerbing, 1991). When examining a construct that exists in the natural world surrounded by other similar constructs, researchers should use indicators that limit the information communicated about orbiting constructs (Colquitt et al., 2019).

Researchers have applied alternate labels to this facet of content validity. Hemphill and Westie (1950) describe "homogeneity of placement" as the extent to which an item applies to a dimension and that it simultaneously does not apply to other dimensions in the description system. Johnston et al. (2014, p. 241) use the label "discriminant content validity" to capture their concern that items assess the intended theoretical constructs and *only* that construct. Colquitt et al. (p. 1) define "definitional distinctiveness" as the degree to which a scale's items correspond more to the target construct than to the definitions of other orbiting constructs. We adopt the label "external congruence" and the synonym "contamination" to unify the various labels in prior literature.

Guion (1977, p. 6) presents relevance (internal congruence) and boundaries (external congruence) as two distinct considerations. An item may be generally relevant to a target construct definition yet still contain contamination from other nearby orbiting constructs. Thus, demonstrating internal congruence does not establish external congruence. Some items can be ambiguous and capture information from multiple neighboring constructs. When an item is meaningfully relevant to a neighboring construct, then contamination remains. When the item is more relevant to the neighboring construct than the target construct, then significant measurement error is systemically introduced into the study data. Having noted that, one can make the case that tests that examine external congruence subsume tests that examine internal congruence. If one removes all contaminants from an item, what remains is relevant only to its intended construct. Despite Guion's view that these facets differ[3], a process that demonstrates the absence of contamination simultaneously demonstrates relevance. Therefore, we present these facets as two nested congruence dimensions. In summary, congruence requires that:

1) Items be highly relevant to their intended construct (internal congruence) and
2) Items lack contamination from other similar constructs (external congruence).

3.2.1 Evaluating Internal Congruence (Relevance)

Researchers have proposed two methods to assess internal congruence: item sorting (also often called Q-sorting, card-sorting or item matching) and item rating. Item sorting exists in several forms. The process involves jurors grouping related items (unbounded by construct definitions) or matching items to provided constructs. Agreement among jurors is calculated as a percent agreement (number of jurors assigning the item to its correct construct divided by the total number of jurors). Researchers have called this metric a factorial validity index (FVI) (Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003), and the proportion of substantive agreement (p_{sa}) (Anderson & Gerbing, 1991). Some researchers have proposed Cohen's kappa as an alternate when using jurors in matched pairs to assess agreement across many decisions

[3] Where we use the label "facet", Guion (1977, p. 5) referred to content validity's dimensions as "conditions".

(Moore & Benbasat, 1991). The unbounded sorting process allows studies to group similar items but says nothing about the relevance of items to a theorized construct. The method neglects the fit between a set of items and a target construct then leaves relevance to the investigator as a matter of judgement. The bounded sorting process recognizes that the items are more relevant to the target construct than other identified constructs but says nothing about that fit's quality. As Shriesheim, Powers, Scandura, Gardiner, and Lankau (1993, p. 395) note: "even though items factor, group, or cluster together, this does not indicate that they are measuring the same theoretical content domains; it only indicates that the items are perceived in a similar manner by respondents". Due to these limitations, item sorting provides only weak evidence of relevance.

Item rating involves jurors assigning a score to rate an item's relevance to a defined construct. One presents definitions of the content domain (the noumenon) to jurors along with each item they need to assess. Davis (1992, p. 196) proposes using the label "relevance" to guide jurors using a Likert scale. Researchers have used several variations of this label, such as "relevant" (Gehlbach & Brinkworth, 2011; Polit, Beck, & Owen, 2007), "representativeness" (Rubio et al., 2003), "correspondence" (Hinkin & Tracy, 1999), "consistent with" (Johnston et al. 2014), "reflects" (Hambleton, 1984; Schriesheim et al., 1993), and "matching our concept" (Colquitt et al., 2019). Gajewski et al. (2012, p.90) demonstrate that, among domain experts, the wording "relevant" encompasses the concept of "correlation". Gajewski et al. (2012) use this wording in their rating scales: 1) item is not relevant, 2) item is somewhat relevant, 3) item is quite relevant, and 4) item is highly relevant.

As with clarity, researchers can calculate agreement on relevance using r_{WG}. A score of 0.70 or above suggests good agreement. Researchers should examine items with poor agreement for rewording and or removal. They may also interview jurors to elicit additional insights to guide rewording.

3.2.2 Evaluating External Congruence (Contamination)

Whereas internal congruence has received intermittent attention in the IS literature, external congruence is largely ignored. The following example demonstrates contamination in IS constructs "validated" for internal congruence. Consider the classic IS constructs perceived usefulness (PU) (Davis, 1989) and relative advantage (RA) (Moore & Benbasat, 1991). Table 3 presents the construct definitions and items involved. Note that these scales share five nearly identical items (shaded). The noumena of interest clearly differ. PU examines the intersection of an information system and job performance, whereas RA examines the relationship between an information system and its precursor. The construct definitions imply additional context. The term "enhance" in the PU definition implies a new system, which suggests that a precursor exists. The RA definition does not require that one uses an innovation in the workplace. However, some RA items clearly imply a workplace setting: does that wording represent an intended referent shift or clutter that reduces the item's clarity? The construct definition could easily apply to an innovation that one uses at home or to part of entertainment at a sporting event.

Table 3. Example of Two Orbiting Constructs

Perceived usefulness: "the degree to which an individual believes that using a particular system would enhance his or her job performance" (Davis, 1989, pp. 320, 340).	**Relative advantage**: "the degree to which an innovation is perceived as being better than its precursor" (Moore & Benbasat, 1991, pp. 195, 216).
Using xxx in my job would enable me to accomplish tasks more quickly.	Using xxx enables me to accomplish tasks more quickly.
Using xxx would improve my job performance	Using xxx improves my job performance
Using xxx would make it easier to do my job.	Using xxx makes it easier to do my job.
Using xxx in my job would increase my productivity.	Using xxx increases my productivity.
Using xxx would enhance my effectiveness on the job.	Using xxx enhances my effectiveness on the job.
I would find xxx useful in my job.	Using xxx gives me greater control over my work.
	The disadvantages of using a xxx far outweigh the advantages.
	Using xxx improves the quality of work I do.
	Overall I find using a xxx to be advantageous in my job.

While examining the definitions one can appreciate the boundaries between the constructs. The complication is that informants answering a survey are generally not aware of the latent construct definitions. When informants consider the survey items, they have only the questions to guide the boundaries they choose to apply. Furthermore, informants can freely interpret each question with different boundaries. Most informants do not understand the principles behind multi-item psychometric scales that contain multiple questions that focus on a single unobservable noumenon. Informants reframe their interpretation anew with each question. The challenge in establishing external congruence involves determining the extent to which informants provide information on the intended (and not some other orbiting) noumenon. Consider the RA question "using *xxx* makes it easier to do my job" in Table 3. Would informants provide information about the relationship between an information system and job performance or between an information system and its precursor? Many individuals that have family "job lists" of household chores may find the label "job" ambiguous. The implied boundary is even more tenuous in organization cultures that emphasize "teamwork" and "family" and that prefer the label "team member" over "employee".

When considering external congruence, one may observe shared properties with the more familiar unidimensionality criterion. However, construct validity tests for unidimensionality that rely on study data are poor indicators of external congruence. Appendix B addresses unidimensionality in detail.

Anderson and Gerbing (1991), Hinkin and Tracey (1999), Rubio et al. (2003), and Johnston et al. (2014) have proposed variations of a multi-step process that combine sorting and rating to evaluate external congruence. First, jurors need to match items with candidate latent constructs, which means researchers need to provide a set of latent construct definitions and a list of items. Jurors review each item and assign it to one latent construct. To make this task manageable, researchers often present definitions in groups of three or four. To maximize the opportunity to expose poor external congruence, researchers should collect plausibly similar and orbiting construct definitions into the same group. In addition, researchers can group new constructs with established venerable constructs, including some that are not in the research model (Colquitt et al., 2019). The list of construct definitions should also include "other" to give jurors the opportunity to expose orbiting constructs that researchers did not anticipate.

Second, jurors need to rate their confidence that the item measures the latent construct. Johnston et al. (2014, p. 250) propose a method for jurors to score their confidence that an item corresponds to the selected construct (from "not at all confident" to "extremely confident"). Alternate anchors include a scale with "item does an extremely bad job" at one end, and "item does an extremely good job" at the other (Colquitt et al., 2019, p. 1265). Researchers should provide A space for comments and suggestions.

The process that Hinkin and Tracy (1999) and Johnston et al. (2014) propose involves each juror rating correspondence for every item against every defined construct (a "fully crossed" design). This approach has two problems (one practical and the other conceptual). From a practical standpoint, rating many items across many constructs constitutes a cognitively demanding exercise (Anderson & Gerbing 1991; Rovinelli et al., 1976). The number of assessment decisions rises quickly as the number of considered items increases. Some researchers report abandoning the process after panel members complained (Hoehle & Venkatesh, 2015). Others suggest dividing scales into subsamples such that each juror assesses a fraction of all items (Colquitt et al., 2019). Of course, such an approach dramatically expands the number of jurors needed for each pre-study iteration. From a conceptual standpoint, when a study informant answers a survey question, the informant picks an answer based on interpreting only that item. A juror who rates an item against multiple construct definitions no longer emulates the main study informant's frame of mind but makes a series of forced judgments; that is, judgments the study informant is neither encouraged nor allowed to make. As a result, the information in those extra judgments no longer represents a main study respondent's perspective.

A nested design constitutes a more representative evaluation. In a nested design, jurors select the one definition that they believe best matches a question, and then rate their confidence the item corresponds to that definition. In this design, jurors have the option to select "other" when they determine that an item aligns closely with some other non-supplied noumenon. This approach has the advantage of being practical for jurors and provides information that conceptually represents the mental frame of informants during the study's main data collection.

Researchers can calculate multiple metrics to evaluate external congruence. Suitable screening metrics include the substantive validity coefficient (c_{sv}) (Anderson & Gerbing, 1991), the index of homogeneity of placement (I_{ij}) (Hemphill & Westie 1950), and the index of item-objective congruence (I_{ik}) (Hambleton,

1984). The substantive validity coefficient (c_{sv}) is suitable to screen for external congruence because it reveals the proportion of informants likely to focus on the intended latent factor when responding to this item. Values above 0.50 suggest adequate congruence, values above 0.61 suggest strong congruence, and values above 0.81 suggest very strong congruence (Colquitt et al. 2019).

The distinctiveness metric htd that Colquitt et al. (2019) propose evaluates congruence more broadly. Whereas the htd metric assumes a fully crossed design, one can adapt a similar calculation to the proposed nested design. We designate that metric htd* (distinctiveness for nested design) and recognize that the scope of data only includes matched item-and-definition pairs. We suggest using the evaluation criteria that Colquitt et al. (2019, p. 1257) propose for *"*weaker average correlations". This criterion provides the most conservative threshold suited for the reduced data set. Values above 0.26 suggest adequate congruence, values above 0.35 suggest strong, congruence and values above 0.48 suggest very strong congruence.

This multi-step matching and rating process addresses both internal congruence and external congruence with a single confirmation metric. Calculating htd is appropriate for fully crossed pre-study designs and htd* is appropriate for nested pre-study designs**.** Researchers should examine items with poor congruence for rewording or removal. Researchers may also interview jurors to elicit additional insight to guide rewording.

3.3 Adequacy

Researchers must "adequately sample" a domain (Guion, 1977, p.7). Each measure should reflect the concept in both content and scope (Johnston et al., 2014). Furthermore, items collectively must capture the full breadth of information about the latent construct (McKenzie, Wood, Kotecki, Clark, & Brey**,** 1999). Adequacy is of greater concern for formative constructs. While reflective items contain redundant information, formative items do not (MacKenzie et al., 2011). Ideally, researchers should collect orthogonal formative indicators. In such a situation, the data that each indicator represents is asymmetric and does not play the same role.

Sufficiency constitutes one aspect of adequacy for formative indices. An index will not accurately reveal an implied formative construct if one omits important facets. This problem of under specifying causal indicators in formative latent constructs appears under various labels in the literature on latent index development. MacKenzie et al. (2011) describe causal indicators as non-redundant measures, which means that they capture different facets of the construct. Removing a measure may omit a unique part of the construct and change its meaning. Petter et al. (2007) observe that causal indicators are non-interchangeable. They note that such indicators need not have similar content and that removing an indicator can alter the construct's conceptual domain. Diamantopoulos (2006) characterizes a formative construct's surplus meaning as the influence of unmeasured causes. We adopt the label "indicator sufficiency" to unify the various labels in prior literature.

A subtle companion consideration is overrepresentation, resulting in unwanted redundancy. Franke, Preacher, and Rigdon (2008) observe that formative indicators have a proportional influence. As a result, adding many measures on a single facet is as detrimental as excluding a facet. The result changes the formative construct's meaning. When building a formative index, researchers should eliminate highly intercorrelated items. Researchers should retain only items that have a distinct influence on the latent variable (Diamantopoulos & Siguaw, 2006). This problem (i.e., over-specifying formative latent constructs) appears under various labels in the literature on latent index development. Haynes et al. (1995) describe "representativeness" as the degree to which elements are proportional to the targeted instrument's facets. They note that content validity is compromised when any facet of a construct disproportionately influences the construct's aggregate score; for example, when a questionnaire contains three items on one facet and only one on another. Diamantopoulos and Siguaw (2006) use the label "parsimony" to discourage redundancy among causal indicators and recommend minimizing multicollinearity. MacKenzie et al. (2011) use the term "unique" and note that each formative indicator should capture a distinct aspect of the causal domain. We adopt the label "indicator parsimony" to unify the various labels in prior literature.

In summary, adequacy requires:

1) Inclusion of all significant facets of the latent factor (indicator sufficiency), and
2) Proportional representation of each facet (indicator parsimony).

3.3.1 Evaluating Indicator Sufficiency

A researcher investigating indicator sufficiency is seeking to expose meaningful facets of a noumenon that have been missed. Demonstrating indicator sufficiency requires one to identify an authority on a latent construct. However, the question arises as to who gets to decide which items one requires and which items one does not. In some situations, domain experts may indeed be authoritative. However, in many situations, domain experts are a poor choice. Two types of formative constructs help understand the challenge: 1) theoretical latent constructs and 2) practical latent constructs.

The scientific community often engages in debates extending over decades to define essential facets of theoretical latent constructs. As a result, the discussion often involves dozens or hundreds of studies that explore the possibilities, then converge on a core set of dimensions. A panel of subject matter experts generally lacks the background needed to effectively select important facets. An ad hoc academic panel lacks the standing to guide the selection until such time as a consensus among scholars emerges.

Practical latent constructs face a challenge of generalizability. IS project success as an example. Whereas substantial literature characterizes IS project success using the "iron triangle" language of cost, functionality and time (Serrador & Turner, 2015), the reality for an individual project depends on the context. Consider a project that Apple Corporation initiated in the late 1990s when the company teetered on the brink of bankruptcy. In that situation, one can easily understand project cost's prominent role in evaluating success. Fast forward 20 years, and Apple sits on one of the largest cash hoards of any corporation at any time in history. In this later situation, readers can easily envision Apple initiating many purely speculative projects whose cost does not represent an important success factor. Only the executives at Apple can decide what they consider important for IS project success. If researchers asked the project's participants (the local domain experts), or academic experts, they are unlikely to collect the important facets. Furthermore, operationalizing a formative construct for one setting is unlikely to generalize to other corporations and settings (Lee & Baskerville, 2003). Such is the plight of practical formative constructs.

Among all facets of content validity, indicator sufficiency is singularly ill-suited for a quantitative evaluation. Therefore, we recommend a qualitative approach. Assuming researchers can identify the authority for defining a formative latent construct, then we recommend that they a conduct a structured interview with such experts. In the interview, researchers can employ an iterative nominal group technique to list, prioritize, and consolidate the important dimensions (McMillan, King, & Tully, 2016). In the absence of an authoritative panel, researchers should gravitate toward reflective scales.

3.3.2 Evaluating Indicator Parsimony

Indicator parsimony concerns whether a scale overrepresents any single facet or domain. Lawshe (1975) proposes a method for subject matter experts to establish that items appropriately sample a formative content domain. In the method, the experts rate each item as: 1) essential, 2) useful but not essential, or 3) not necessary. Lawshe (1975) calls this metric the content validity ratio (CVR). The metric is a linear transformation of the ratio of the number of jurors who rate an item as essential to the total number of jurors on the panel.

Despite its prominence in the content validity literature, CVR is a poor choice for several reasons. First, the metric is characterized as an "overall" content validity metric. CVR instruments collect data using the anchor "essential", thereby limiting applicability to indicator parsimony. This anchor has poor relevance for the remaining content validity dimensions clarity, congruence, indicator sufficiency, or dependability. Second, in order to serve as a method to assess indicator parsimony, researchers need to provide additional instructions to jurors so that they understand that the purpose the anchor "essential" is to parsimoniously represent each facet of a formative indicator. The literature lacks this nuanced instruction. If a panel does not understand this constraint, then one can expect a high type I error as jurors assess multiple items for a single facet as "essential". In this scenario, redundant items remain in the scale and overrepresent one or more dimensions. Third, in situations where panel jurors understand what researchers intend by "essential", the content validity ratio has high type II error. When evaluating three or more indicators as candidates for a single trait, CVR will likely reject all three. An example illustrates this problem.

Consider a hypothetical formative latent construct to measure IS project success. Figure 3 shows a set of candidate survey items that a well-intentioned research team might brainstorm to measure IS project success. The researchers present these indicators to jurors to rate using the CVR scale of "essential". If

jurors do not recognize the four distinct traits, then one or more traits is unmarked and omitted in the final scale. Assuming the panel understands what researchers intend by “essential”, each juror should mark one item for features, one item for cost, one item for time, and one item for quality. In each domain, a uniform distribution will exist if the jurors judge the items as equally suitable. Thus, each item in the cost domain will receive a 33.3 percent rating as “essential”, which drives CVR for every cost indicator into the “reject as non-essential” category. Researchers could plausibly use any of the three cost indicators to capture the cost dimension of IS project success, but CVR rejects all three. Depending on the number of jurors who rate items for this scale, a similar problem would occur for time and quality if the jurors judge the two items as equally suitable.

We propose redefining the process of selecting parsimonious formative indicators into a hierarchical problem. Rather than presenting a definition of the latent construct to the review panel, researchers can formally define each facet (e.g., features, cost, time, and quality as Figure 4 depicts). Jurors then follow the external congruence process to match items to facet definitions and rate their confidence in that match. After scoring, researchers would retain items with the best htd* for each dimension. The result of this process is an evidence-based evaluation of a “best”-fitting formative indicator that eliminates redundant formative indicators. In situations where two items associated with the same facet have equivalent congruence, researchers can use scores for relevance as a guide to select the best indicator. Researchers would produce a parsimonious set of formative indicators such as the example in Figure 5.

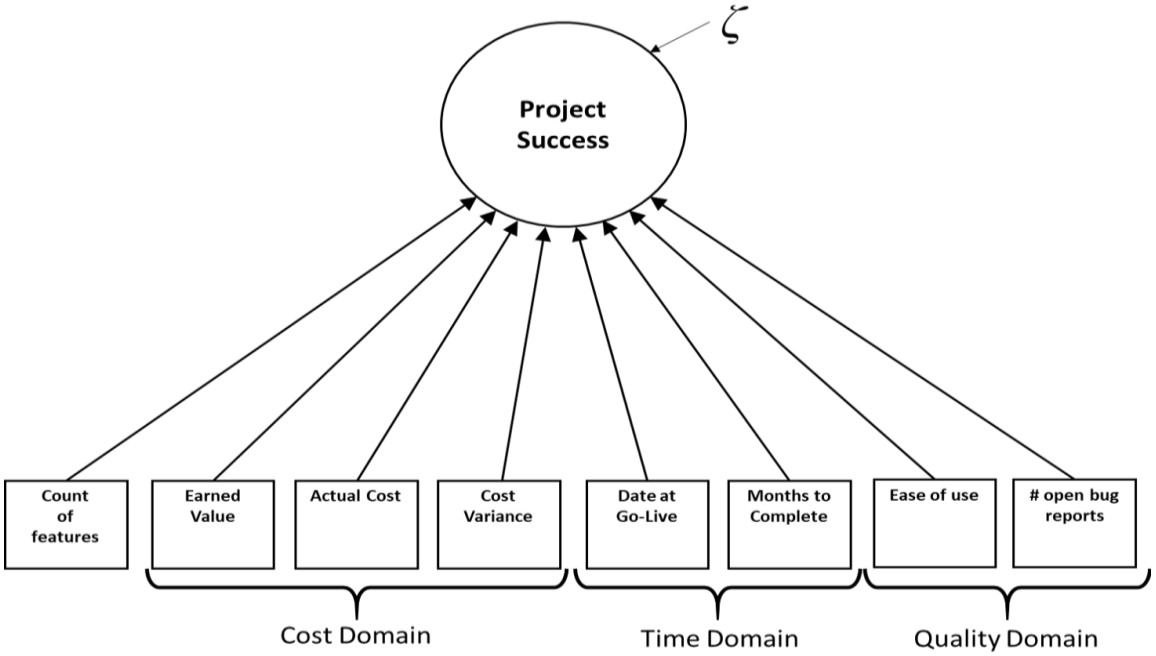

Figure 3. Candidate Formative Indicators

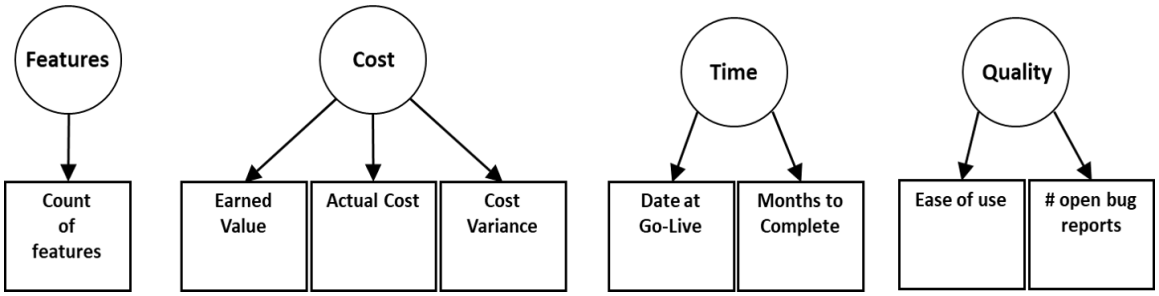

Figure 4. Formative Construct Modeled as Multidimensional

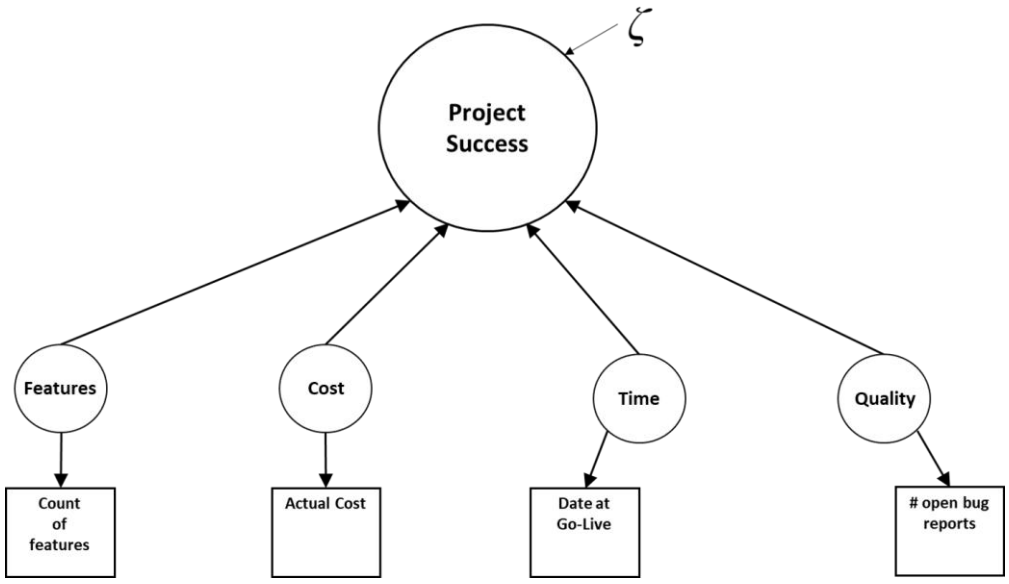

Figure 5. Parsimonious Formative Indicators

In situations where dimensions of a formative latent construct are not directly observable, researchers can improve measurement by modeling a second-order formative construct then measure each dimension with its own reflective scale (Wright, Campbell, Thatcher, & Roberts, 2012). Indicator sufficiency becomes domain sufficiency and remains as a content validity consideration. However, overrepresenting items (indicator parsimony) no longer poses a concern because this is managed within the calculation dynamics of the first order reflective construct.

3.4 Dependability

A scale should be a dependable measure of its target noumenon for the intended population. Dependability "embraces elements of both the stability implied by the rationalistic term *reliable,* and the tractability required by explainable changes in instrumentation" (Guba, 1981, p. 81, emphasis in original). The concept reliability emerges from "assumptions of repeatability, replicability and consistency" (Baskerville, Kaul, & Storey, 2017, p.4). According to Guion (1977, p. 7), "content must be reliably observed and evaluated". He emphasizes "standardization that allows at least some assurance that the stimulus content is presented in the same way for all examinees and the response content is evaluated according to the same rules by all observers" (p. 6). Two aspects of dependability are important in the content domain: 1) consistent presentation and 2) standard interpretation.

In discussing content clarity in Section 3.1, we focus on projecting meaning from the construct definition to individual items. However, a single item does not present or capture a noumenon's content domain. While clarity, congruence, and adequacy examine quality at the item level, dependability requires a high-quality of the scale as a whole. The information that a scale collects in aggregate captures the implied noumenon. Therefore, consistent presentation is a function of using multiple items to measure a latent construct. This requirement of a "census of all concepts that form the construct" (Jarvis, MacKenzie, & Podsakoff, 2003, p. 202) applies to formative indices and reflective scales. According to Fitzpatrick (1983, p. 10), "poor items that appear in a test will jeopardize the fit between the test and its definition and, therefore, the reproducibility of any test results that are obtained". Dependability requires researchers to consistently *present* content for the scale as a collective.

Presentation guides interpretation. Dependable evaluation emerges from a stable interpretation of the scale. As Lennon (1956, p. 296) notes, "appraisal[s] of *content* validity must take into account, not only the content of the questions, but also the *process* presumably employed by the subject in arriving at his response" (emphasis in original). Dependability deals with whether the interpretation remains stable for informants in the target population. An item that one juror panel judges as relevant should remain relevant for another panel. An item that a panel judges as clear one week should remain clear to that same panel three or four weeks later. Each content validity assessment should remain demonstrably consistent for different panels or the same panel at different times (Parker, Handson, & Hunsley, 1988).

Inferences of consistent presentation and stable interpretation have their foundation in pre-study methodology. Careful selection of jurors allows the lens of investigation used in the pre-study to emulate informants in the main study. When jurors emulate the evaluation processes of main study informants, then consistent and reliable feedback for clarity, congruence, and adequacy also support an inference of standard and dependable interpretation. Many scholars recommend using subject matter experts who have domain knowledge and are representative of the population of interest (Lawshe, 1975; Anderson & Gerbing, 1991; McKenzie et al., 1999; Rubio et al., 2003). As observed by Guion (1977, p.7) "the content domain of a job is better judged by people who have performed that job, or supervised its performance, or have done careful analysis of it, than by those whose main qualifications are degrees".

In summary, dependability requires:

1) The scale as a collective should consistently present a noumenon's content to jurors whom researchers draw from the target population (content consistency), and
2) The scale as a collective should be interpreted in stable manner by Jurors drawn from the target population (content stability).

Beyond carefully selecting the jury panel, researchers can collect supporting evidence of standard and reliable interpretation via various means. Goodwin and Leach (2003) suggest that researchers use interviews with jurors so that they can analyze the jurors' responses to items. Goodwin and Leach also suggest that researchers systematically examine similarities and differences across various subgroups. Finally, they suggest that researchers study changes in responses over time, such as with a pre-study assessment test-retest. Supporting evidence can bolster confidence in content dependability. For example:

1) Qualitative evidence from jurors (interviews) describing how they interpret items supports overall claims of dependability and provides suggestions for improvement when needed.
2) Evidence of internal consistency reliability using main study data indirectly supports content consistency. Direct evidence of content consistency is obtained by aggregating juror ratings for relevance and confidence across all items in a scale.
3) Evidence that the scale remains stable in a juror assessment test-retest supports content stability.

3.4.1 Juror Selection

The power of each well-chosen jurist is equivalent to data from at least 10 main-study informants (Gajewski et al., 2012), which justifies great care in the selection of appropriate jurors for pre-study panels. Studies using panels for pre-testing scales report using academic experts, domain experts, and even naïve jurists (Colquitt et al., 2019; Johnston et al., 2014; Robinson, 2018; Tan, Benbasat, & Cenfetelli, 2013). Researchers recruit naïve jurists from an extended population of people available to answer surveys often via the Internet or brokers such as MTurk. Naïve jurors are suitable when they are similar to the target population. However, they are ill suited when the target population possesses context-specific vocabulary or experience that guide a local interpretation and reaction to survey questions. Furthermore, naïve jurists generally lack the background that one needs to understand the scientific jargon in formal construct definitions.

Academic expert panels can provide researchers with jurors who understand scientific jargon and behavioral science research methods, such as psychometric testing. As such, academic experts are well-positioned to judge the clarity of a formal construct definition and the degree to which the formal definition projects into the operational definition of specific survey items. In addition, academic experts understand why many scales have redundant items and can appreciate a high-quality item's characteristics (e.g., avoiding double-barreled questions and multiple negations).

Unfortunately, academic experts also bring familiarity with established scales and may use this background to a priori align items they recognize to known constructs without applying the cognitive effort that they need to make an independent assessment (Colquitt et al., 2019). Furthermore, academic experts rarely represent a study's target population. Academics use a different vocabulary. They may view what is clear differently or interpret boundaries for definitions and concepts in a different way than a subject matter expert who lives and functions in the target domain. In general, researchers should exercise

caution when including students and academics as jurors in pre-studies unless they represent the target population (MacKenzie et al., 2011).

Practitioner (domain) experts (also called subject matter experts) are valuable because they emulate study informants and plausibly interpret items in the same manner as a study informant. Researchers need to take care when providing instructions and additional information to domain experts because that information may establish a mental frame that no longer represents the target population. For example, study informants generally lack awareness of the proposed theoretical model. Therefore, domain expert jurors represent an appropriate choice for assessing item clarity but may face challenges in assessing some aspects of construct clarity. We recommend that researchers assess latent construct definition clarity with both academic and domain jurors and use domain experts to assess item level clarity. Domain experts represent a good choice to assess all other content validity aspects.

3.4.2 Juror Interviews

Juror interviews during pre-studies focus on the "fit for purpose" of survey items to measure a target noumenon. Conducting interviews prior to a matching and rating exercise will often introduce information that could alter a juror's ability to represent the mental frame of main study informants. As a result, interviews provide the most value after jurors have examined and rated items using the pre-study feedback instrument.

The process that jurors use to understand each item individually is important. Informants in the final study can freely evaluate each question independently, so jurors should describe how they evaluated each item on its own merits. The process a juror follows to interpret an item may take the form of a story or a logical step-by-step progression. The interpretations that jurors describe reveal the relative importance and value of specific words and phrases. However, a consistent conclusion has more importance than a consistent story or logical path. Ultimately, main study informants present a judgement or opinion on the noumena they conclude a study investigates. Dependability examines which noumena from among a universe of latent constructs jurors have chosen. Subtle differences among jurors that reveal alternate noumena from a constellation of orbiting constructs can provide essential information for researchers to adjust survey items. Individual jurors that equivocate between two or three possible noumena can similarly reveal content domain problems.

In addition to exposing different interpretations among jurors and uncertainty in individual jurors, researchers can pursue suggestions from jurors for vocabulary and phrases that bring them to the target noumenon quicker and more consistently. It is a rare researcher who has the same use of vocabulary as a subject matter expert from the population under investigation.

3.4.3 Evaluating Content Consistency

Researchers familiar with reliability metrics applied to study data will recognize internal consistency assessments. Popular tests include Pearson's correlation coefficient, Cronbach's coefficient alpha, and/or Dillon-Goldstein's rho (DeVellis, 2012). These statistics infer reliability by demonstrating strong correlation among all items in a reflective scale or across multiple uses of an instrument. High inter-item correlations suggest that reflective items share a common cause. Although such correlations do not provide insights into what that "same thing" is, an important practical aspect of dependability involves establishing that multiple measures capture the same thing. However, using study data to demonstrate content consistency presents several problems:

- Since researchers do not expect high inter-item correlations for causal indicators, they are not suitable to assess formative instrument reliability[4].
- High inter-item correlations may not result from a common cause (the latent factor) but rather individual items affecting each other. Traditional internal consistency metrics assume a common cause and offer no ability to expose the latter (DeVellis, 2012).
- Alpha faces particular challenges as a metric for reliability and internal consistency as Sijtsma (2008, p. 114) notes:

[4] Straub et al. (2004, p. 400) abandon the need to assess reliability for formative constructs: "It is not clear, therefore, that reliability is a concept that applies well to formative constructs".

> *Its value depends only on the sum of interitem covariances. Thus, all that alpha can reveal about the "interrelatedness of the items" is their average degree of "interrelatedness", provided there are no negative covariance, and keeping in mind that alpha also depends on the number of items in the test. …This says very little if anything about internal consistency…. Alpha "only" is a lower bound to reliability and not even a realistic one.*

- Alpha requires all items to have equal importance in measuring a latent variable. This implies that all items have equivalent variations and the same degree of precision such that all true loading scores are equal (called tau-equivalence): $Var(\lambda_i) = Var(\lambda_j)$ for all i and j. This assumption is rarely met in practice (Cho & Kim, 2015).

Coefficient omega (ω) is a good choice when using covariance based structural equation modeling to assess a measurement model (McDonald, 1999, p. 59). Coefficient omega (ω) because it does not assume tau-equivalence and one can calculate it for unidimensional and multidimensional constructs (Cho & Kim, 2015). The unidimensional version of omega (ω_u) is a good choice when using PLS methods to assess a measurement model. Many statistics packages report this statistic as composite reliability (Cho & Kim, 2015; Hair, Sarstedt, Ringle, & Mena, 2012).

Despite their popularity, these study-data approaches do not focus the informant's attention to the context in which they interpret these items. That is, study subjects provide information on their experience (the intersection of the implied latent factor and their experience.) This focus allows researchers to only indirectly measure content consistency and, therefore, provides only weak support for dependability. Dependability is more general and subsumes the concept of internal consistency reliability (Wiliam, 1993). Guion (1997) refers to reliability in the sense of stability, consistent meaning, and standard interpretation, not internal consistency reliability[5]. Internal consistency *supports* a claim that a scale reliably measures a common "something" but not that the noumenon of interest is consistently that "something".

Instead, content consistency assesses dependable presentation at the intersection of the real latent factor and the survey scale. A metric that aggregates juror ratings at the scale level supports a claim of content consistency. Researchers can aggregate juror responses on relevance, and separately confidence, at the scale level to provide suitable evidence. Polit et al. (2007) suggest calculating a scale-level CVI metric by averaging the item-level metrics. This approach inherits CVI's weaknesses (dichotomizing data from a four-item rating scale). Using the multi-item agreement index $r^*_{WG(J)}$ to calculate relevance, and separately confidence, allows researchers to assess content consistency for a scale during a pre-study. A score of 0.70 or above suggests agreement. Researchers should examine scales with poor agreement for rewording or item removal.

3.4.4 Evaluating Content Stability

Researchers most commonly evaluate stability by comparing data collected at two different points in time (Kimberlin & Winterstein, 2008). IS scholars who use repeated full-study samples with a finalized instrument have employed this method (Hendrickson, Massey, & Cronan, 1993; Torkzadeh & Doll, 1994). As we mention in Section 1, iterative tests and instrument revision are often impractical when using final study data. Some psychometric dynamics also make study data test-retests a poor fit for establishing content stability. Final study informants provide information at the intersection of the implied latent construct and their experience. For many latent constructs, this intersection is a moving target. For example, some latent constructs represent traits (e.g., personal innovativeness with IT) that are relatively stable, whereas others involve states (e.g., perceived ease of use) that change with the setting and exposure. One faces problems establishing stability for a latent construct when using study data because training, experience, or any number of other factors may alter the way informants perceive their current state (e.g., ease of use) between test and retest. A stable scale will appropriately report a change in state, but this change will simultaneously fail the test-retest assessment.

The relationship between an item and its latent construct should be stable even when the latent construct describes a state that changes[6]. Therefore, researchers can suitably apply the test-retest method during the pre-study for all forms of survey instruments. The test-retest method is particularly attractive due to the iterative nature of content validity pre-studies. Each time researchers adjust a survey instrument, they

[5] "This does not refer to internal consistency" (Guion, 1977, p. 7)

[6] An exception arises when a noumenon's definition changes due to advances in theory or changes to the environment. We address this point in Section 6. Such changes provide one trigger that necessitates that researchers revalidate existing scales.

should assemble jury panels, gather feedback again, and reassess all forms of content validity. Eventually, a stable instrument emerges, with strong correlation between scores from successive panels.

The test-retest procedure involves administering the pre-study feedback instrument to a panel of jurors, and then administering the instrument again at a later date (Kimberlin & Winterstein, 2008). A four-week interval between test and retest suits scale-development purposes, though, in some cases, researchers may be able to support intervals as short as two-and-a-half hours (Shaft, Sharfman, & Wu, 2004). Researchers calculate ICC(K) from two feedback sessions to simultaneously evaluate absolute consensus and relative consistency when the retest involves new jurors. Alternately, researchers should calculate ICC(A,K) when the retest uses the same jurors. A score of 0.70 or above demonstrates stable scales. Researches should revise scales with poor ratings and repeat the process.

4 Methods and Metrics

Jurors analyze and assess the items themselves and provide information at the intersection of a scale and a noumenon. Their qualification is their ability to interpret the noumenon in the same way as a final study participant. Information that researchers collect in a pre-study involves characteristics such as the item's clarity, its relevance to the noumenon, and the degree to which it corresponds with one noumenon or another.

Whereas conclusion validity uses correlation and covariation techniques to study the relationships between latent constructs, the pre-study focuses on establishing "fit for purpose" of survey items. Therefore, statistical techniques to assess agreement among jurors represent the primary tools to analyze content validity. Two families of metrics are candidates for this analysis: consistency and consensus estimates. Each addresses different alignment properties across responses. Consistency estimates measure reliability (researchers sometimes call these metrics inter-rater reliability (IRR)) and provide information on the jurors ("are the jurors alike?"). Consensus estimates measure agreement (researchers sometimes call these metrics inter-rater agreement (IRA)) and provide information on the target instrument ("does the item/instrument have a desired property or characteristic?"). A more complete accounting of differences and suitability is available elsewhere (LeBreton, Burgess, Kaiser, Atchley, & James**,** 2003; LeBreton & Senter, 2008; Stemler, 2004). Established conventions recognize that consensus estimates from the inter-rater agreement family of metrics are well suited for assessing juror data (Kozlowski & Hattrup, 1992).

4.1 Agreement Metrics for Ordinal data

Table 4 details suitable agreement metrics to test ordinal scores such as those recommended for clarity and relevance. The table provides information on the calculation mechanics and evaluation criteria. Early attempts to assess content validity involve the proportion-of-agreement metrics CVR, iCVI, and kappa*, which dichotomize responses from a three- or four-choice scale into the categories acceptable and unacceptable. This process of converting ordinal scores into categorical data discards the granular information from each step in the scale. Kappa* includes a correction for chance agreement and, thus, addresses a weakness in iCVI. CVR uses a critical values table that includes correction for chance agreement but is tuned to data collected on a three-anchor scale specific to "essential".

The correspondence index htc constitutes an alternate metric that retains information across the full range of option choices. This calculation divides the average rating by the number options in the scale and, thereby, creates an index where the value 1 represents perfect agreement. This metric is general purpose in that it suits scales with more than four options. However, htc does not include a correction for chance agreement.

The within-group inter-rater agreement[7] metric r_{WG} constitutes a general-purpose statistic that does include correction for change agreement. This statistics is suitable for pre-studies of psychometric scales, with the advantage of established evaluation criteria. Like htc, r_{WG} accommodates rating scales with more than four options. Higher granularity scales are appropriate when jurors can make distinctions across the full range. We recommend using r_{WG} to test shared perceptions of clarity and relevance.

[7] Although James, Demaree, and Wolf (1984) originally proposed r_{WG} as a reliability/consistency metric, subsequent critiques have firmly placed it in the agreement/consensus domain (Kozlowski & Hattrup, 1992).

Table 4. Agreement Metrics Applicable to Ordinal Data

Metric	Details	
r_{WG} Recommended test for clarity and relevance	**Within-group inter-rater agreement** (James et al., 1984)	
	$r_{WG} = 1 - \left(\frac{s_x^2}{\sigma_{EU}^2}\right)$	s_x^2 = variance of juror's ratings σ_{EU}^2 = variance of uniformly distributed error
	Correction for chance agreement: $\sigma_{EU}^2 = \frac{(A^2-1)}{12}$	A = number of response alternatives (size of scale). EU uniform distribution is justifiable for pilot studies (Brown & Hauenstein 2005)
	Evaluation criteria (LeBreton & Senter 2008) Very strong ≥ 0.91 Strong 0.71 to 0.90 Justifies aggregation* ≥ **0.70** Moderate 0.51 to 0.70 Weak 0.31 to 0.50 Lack ≤ 0.30	*Values suitable to justify aggregation vary depending on application. LeBreton & Senter suggest that 0.70 may be acceptable for newly developed measures but may be too low for established measures.
htc	**Correspondence index** (Colquitt et al., 2019)	
	$htc = \frac{\left(\frac{\sum rs}{j}\right)}{a}$	rs = rating score j = number of jurors a = number of anchors in scale
	No correction for chance agreement	
	Evaluation criteria (Colquitt et al., 2019) Very strong ≥ 0.91 Strong 0.87 to 0.90 Moderate **0.84** to 0.86 Weak 0.60 to 0.83 Lack of ≤ 0.59	Criteria tables for additional interpretational categories of Colquitt et al. (2019)
CVR	**Content validity ratio** (Lawshe, 1975)	
	$CVR = \frac{n_e - \left(\frac{N}{2}\right)}{\left(\frac{N}{2}\right)}$	n_e = number of jurors rating item as "essential" N = number of jurors
	Correction for chance agreement imbedded in critical value table Assume jurors rate on three-point scale for "essential" Scores are dichotomized (∴ some loss of information)	
	Evaluation criteria (Wilson, Pan, & Schumsky, 2012) 5 jurors ≥ 0.736 7 jurors ≥ **0.622** 10 jurors ≥ 0.520	Criteria values tables for 5 ≤ N ≤ 40 Wilson et al. (2019)
iCVI	**Content validity index** (item level) (Davis, 1992; Polit et al., 2007)	
	$iCVI = \frac{n_r}{N}$	n_r = count of jurors selecting relevance scores 3 and 4 N = total number of juror
	No correction for chance agreement. Assume jurors rate on four-point scale Scores are dichotomized (∴ some loss of information	
	Evaluation criteria (Davis, 1992) Good ≥ 0.80 Moderate **0.70** to 0.80 Unacceptable ≤ 0.70	Critical value is dependent on number of jurors. iCVI > 0.78 is excellent regardless of the number of jurors (Polit et al., 2007

Table 4. Agreement Metrics Applicable to Ordinal Data

Kappa*	**Modified kappa statistic** (Polit et al., 2007)	
	$\kappa^* = \frac{(iCVI - p_c)}{(1-p_c)}$	iCVI = Content validity index p_c = probability of chance agreement
	Correction for chance agreement: $p_c = \left[\frac{N!}{(A!(N-A)!)}\right] 0.5^N$	N = number of jurors A = number agreeing on good relevance (3 or 4 on four-point scale)
	Evaluation criteria (Polit et al., 2007) Excellent ≥ 0.74 Good **0.60** to 0.74 Fair 0.40 to 0.59	Criteria value tables for 3 ≤ N ≤ 9 Polit et al. (2019)
I_{ik}	**Index of item-objective congruence** (fully crossed design) (Hambleton, 1984)	
	$I_{ik} = \frac{(N-1)\sum_{n=1}^{n} X_{ijk} - \sum_{i=1}^{N}\sum_{j=1}^{n} X_{ijk} + \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$	I_{ik} = Index for item **k** on construct **i** N = number of constructs (i=1..N) n = Number of jurors (j = 1…n) X_{ijk} = rating (-1,0,+1) of item **k** as a measure of construct ***i*** by juror ***j***.
	No correction for chance agreement	
	Evaluation criteria not provided; judging statistic values "is best done after some experience is gained with content specialists' ratings and with the index itself" (Hambleton, 1984, p. 221)	

The final metric in the table, the index of item-objective congruence (I_{ik}), uses a fully crossed rating design to assess congruence. It evaluates relevance and includes a penalty for contamination; as such, it works well as an overall congruence metric. However, due to the lack of evaluation criteria and the burden of a fully crossed pre-study design, we recommend other methods and metrics.

4.2 Agreement Metrics for Nominal/Categorical data

The process we propose for evaluating congruence includes a matching exercise in which one identifies the best-fitting definition (or "other" when none prove suitable) for each item. Table 5 details suitable agreement metrics that researchers can use to assess nominal (categorical) matches in situations without an implied order among the choices. The proportion of substantive agreement metric (p_{sa}) calculates a simple proportion of agreement, which makes it suitable as a screening test for internal congruence. The substantive validity coefficient (c_{sv}) has the advantage that it includes an explicit penalty when jurors match an item to an orbiting construct definition. This penalty integrates contamination into the calculation, which makes the metric suitable for screening external congruence.

We describe these metrics as "screening tests" because they establish a necessary precondition of congruence but are not sufficient to disprove contamination. Grouping related items fails to assess the quality of correspondence to an intended noumenon or the degree of contamination from an orbiting noumenon.

Table 5. Agreement Metrics Applicable to Categorical Data

Metric	Details	
c_{sv} Recommended as screening test for overall congruence	**Substantive validity coefficient** (Anderson & Gerbing, 1991)	
	$c_{sv} = \frac{(n_c - n_o)}{N}$	n_c = number of jurors who matched the item correctly n_o = number of jurors who matched to an orbiting construct N = total number of jurors
	No correction for chance agreement	

Table 5. Agreement Metrics Applicable to Categorical Data

	Evaluation criteria (Colquitt et al., 2019) Very strong ≥ 0.81 Strong 0.61 to 0.80 Moderate **0.51** to 0.60 Weak 0.05 to 0.50 Lack of ≤ 0.04	Criteria value tables for additional interpretational categories provided by Colquitt et al. (2019)
p_{sa}	**Proportion of substantive agreement** (Anderson & Gerbing, 1991) Factorial validity index (Rubio et al., 2003)	
	$p_{sa} = \frac{n_c}{N}$	n_c = number of jurors who matched the item correctly N = total number of jurors
	No correction for chance agreement.	
	Evaluation criteria (Colquitt et al., 2019) Very strong ≥ 0.91 Strong 0.82 to 0.91 Moderate **0.72** to 0.81 Weak 0.39 to 0.71 Lack of ≤ 0.38	Criteria value tables for additional interpretational categories provided by Colquitt et al. (2019)

4.3 Agreement Metrics Integrating Categorical and Ordinal data

Assessing overall congruence requires statistics that integrate both categorical (matching) and ordinal (rating) data. Table 6 identifies suitable metrics to test congruence. We recommend that researchers use the htd metric to assess congruence when they can collect correspondence ratings for all possible combinations of items and construct definitions (a fully crossed design). However, given the fully crossed design's high cognitive burden and the likelihood that fatigue will lead to an incomplete dataset, researchers may often choose a nested design in which they collect correspondence judgments only for the item-definition match. We recommend that researchers use the htd* statistic when using the nested design.

Table 6. Agreement Metrics Integrating Categorical and Ordinal Data

Metric	Details	
htd Recommended test for congruence (fully crossed design)	**Distinctiveness** (fully crossed design) (Colquitt et al., 2019)	
	$\text{htd} = \frac{\left(\frac{\sum icr - \sum ocr}{j}\right)}{(a-1)}$	icr = correspondence to intended factor ocr = correspondence to orbiting factor j = number of number of jurors a = number of anchors in scale
	No correction for chance agreement Correspondence ratings obtained for intended construct and all presented orbiting constructs (fully crossed design).	
	Evaluation criteria (Colquitt et al. 2019) Very strong ≥ 0.35 Strong 0.27 to 0.34 Moderate **0.18** to 0.26 Weak 0.04 to 0.17 Lack of ≤ 0.03	Criteria tables for additional interpretational categories provided by Colquitt et al. (2019)

Table 6. Agreement Metrics Integrating Categorical and Ordinal Data

<table>
<tr><td rowspan="4">htd*

Recommended test for congruence (nested design) and indicator parsimony</td><td colspan="2">Distinctiveness* (nested designs) (NEW)</td></tr>
<tr><td>htd* = $\frac{\left(\frac{\sum icr^* - \sum ocr^*}{j}\right)}{A}$</td><td>icr* = correspondence to matched intended factor
ocr* = correspondence to matched orbiting factor
j = number of jurors
A = maximum anchor value on correspondence scale</td></tr>
<tr><td colspan="2">No correction for chance agreement
*Correspondence ratings obtained for matched items & constructs (nested design).</td></tr>
<tr><td>Evaluation criteria (Colquitt et al. 2019)
Very strong ≥ 0.48
Strong 0.35 to 0.47
Moderate 0.26 to 0.34
Weak 0.12 to 0.25
Lack of ≤ 0.11</td><td>Mathematically equivalent to htd but with restricted data collection.
Criteria from the most conservative htd table “weaker average correlation” (Colquitt et al., 2019).</td></tr>
</table>

4.4 Scale-level Agreement Metrics

Testing dependability in a pre-study includes determining that all items collectively represent the target construct. Table 7 identifies suitable metrics to test scales as a collective. The scale-level content validity index (sCVI) averages item-level metrics. The underlying calculations are based on a four-choice ordinal scale that researchers dichotomized for analysis. The result discards data, which may help inform decisions. Some researchers have proposed the evaluation criteria of 0.80 without explanation.

The multi-item agreement index $r^*_{WG(J)}$, a general-purpose calculation, works with all size scales. In addition, detailed analysis provides evaluation criteria (LeBreton & Senter, 2008) to guide the scale-development process. Researchers should use the $r^*_{WG(J)}$ metric when assessing content consistency.

Table 7. Scale-level Agreement Metrics Applicable to Content Validity

<table>
<tr><th>Metric</th><th colspan="2">Details</th></tr>
<tr><td rowspan="4">$r^*_{WG(J)}$

Recommended test for content consistency using scores for relevance and confidence.</td><td colspan="2">Multi-item agreement index (Lindell & Brandt, 1999)</td></tr>
<tr><td>$r^*_{WG(J)} = 1 - \left(\frac{\bar{s}^2_x}{\sigma^2_{EU}}\right)$</td><td>$\bar{s}^2_x$ = average variance of juror ratings
σ^2_{EU} = variance of uniformly distributed error</td></tr>
<tr><td>Correction for chance agreement:
$\sigma^2_{EU} = \frac{(A^2 - 1)}{12}$</td><td>A = number of response alternatives
EU uniform distribution is justified for pilot studies (Brown & Hauenstein, 2005)</td></tr>
<tr><td>Evaluation criteria (LeBreton & Senter, 2008)
Very strong ≥ 0.91
Strong 0.71 to 0.90
Justifies aggregation ≥ 0.70
Moderate 0.51 to 0.70
Weak 0.31 to 0.50
Lack ≤ 0.30</td><td>Mathematically equivalent to r_{WG} when j = 1, ∴ cutoff values derived by LeBreton & Senter (2008) also applicable to this scale-level index</td></tr>
<tr><td rowspan="4">sCVI(A)</td><td colspan="2">Scale content validity index (average agreement) (Davis, 1992; Polit et al., 2007)</td></tr>
<tr><td>sCVI = $\frac{\sum iCVI}{k}$</td><td>iCVI = content validity Index (item level)
k = number of items</td></tr>
<tr><td colspan="2">Use kappa* instead of iCVI to correct for chance agreement</td></tr>
<tr><td>Evaluation criteria (Davis, 1992)
Very strong ≥ 0.80</td><td></td></tr>
</table>

4.5 Test-retest Metrics

Researchers can use test-retest methods to demonstrate content stability. Table 8 identifies metrics for testing stability across two panel events. To assess test-retest stability, researchers most commonly demonstrate correlation across successive tests. The test-retest reliability coefficient is simply the Pearson product-moment correlation coefficient applied to test-retest scores. Researchers have generally accepted this metric to establish inter-rater reliability. The interclass correlation (ICC) statistic is an alternative. ICC provides information on both inter-rater agreement and inter-rater reliability. Researchers have established many ICC variations to address different agreement scenarios (McGraw & Wong, 1996). When evaluating the stability of raters (the jurors), researchers should use the ICC(1) formula (LeBreton & Senter, 2008). When evaluating a survey instrument's stability across two measurement events, the survey instrument is the evaluation target. In this situation, researchers should use the ICC(K) formula (LeBreton & Senter, 2008). This metric appears in some literature with the label ICC(2) or ICC(1,K). In situations where researchers convene a new set of jurors for the retest event, then they should calculate the statistic ICC(K) based on a one-way random-effects ANOVA. Alternately, if researchers convene the same juror panel for the retest event, then they should calculate the statistic ICC(A,K) based on a two-way mixed-effects ANOVA (LeBreton & Stenter, 2008).

Table 8. Test-retest Metrics for Content Stability

Metric	Details	
ICC(K) Recommended for test-retest stability	**Interclass correlation**, ICC(K) (McGraw & Wong, 1996; LeBreton & Senter, 2008)	
	$x_{ij} = \mu + r_i + w_{ij}$ (1) $ICC(K) = \frac{MS_R - MS_W}{MS_R}$ (2) $ICC(A,K) = \frac{MS_R - MS_W}{MS_R + \frac{MS_C - MS_E}{N}}$ (3)	x_{ij} = ANOVA μ = overall mean r_i = difference from mean for ith trial j = 1..k = number of jurors (columns) i =1..n = number of items rated (rows) w_{ij} = residual effects and error MS_R = mean square for rows MS_W = mean square within (residual sources of variance) MS_C = mean square of columns (jurors) MS_E = mean square error N = number of items rated
	Spearman-Brown correction (LeBreton & Senter, 2008) MS values derived from ANOVA Use formula (2) with one-way random effect ANOVA if different set of jurors for retest Use formula (3) with two-way mixed effects ANOVA if same set of jurors for retest	
	Evaluation criteria (LeBreton & Senter, 2008) Justifies aggregation **0.70** to 0.85	
r_{xy}	**Pearson's correlation coefficient**, sample	
	$r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x s_y}$	$\bar{x}$ = average of scores from test (trial 1) $\bar{y}$ = average of scores from retest (trial 2) n = number of jurors s_x = standard deviation for trial 1 s_y = standard deviation for trial 2
	Evaluation criteria (Nunnally, 1978; Shrout & Fleiss, 1979) Very good ≥ 0.80 Acceptable **0.75** to 0.80	

4.6 Summary of Content Validity Literature

Table 9 summarizes the literature that provides advice for examining and testing content validity. As one can see, only Guion (1997) identifies all content validity's facets (designated in the table with the letter "D" for discussion). Various scholars propose methods (designated with the letter "P" for process) to examine specific aspects of content validity, although, in many cases, they leave evaluation as an exercise in subjective judgement. Explicit recommendations for quantitative tests (designated with the letter "T" for test) are less prominent.

Table 9. Content Validity Recommendations

Source	Clarity		Congruence		Adequacy		Dependability		Lit domain
	Construct clarity	Item clarity	Internal congruence (relevance)	External congruence (contamination)	Indicator sufficiency	Indicator parsimony	Content consistency	Content stability	
Anderson & Gerbing (1991)	D	D	D,P,T P_{sa}	D,P,T C_{sv}	D			D	Social science (psychology)
Colquitt et al. (2019)			D,P,T P_{sa}, htc	D,P,T C_{sv}, htd					Social science (psychology)
Davis (1992)		D	D,P,T CVI		D				Healthcare (nursing)
DeVillis (2012)	D	D,P	D,P	D,P	D,P		D,P,T α	D,P,T r_{xy}	Social science (methods)
Diamantopoulos & Siguaw (2006)					D	D			Social science (management)
Drost (2011)			D				D,P,T α	D,P,T r_{xy}	Education
Fitzpatrick (1983)	D		D		D		D		Social science (methods)
Gajewski et al. (2012)			D,P,T CVI,CVR						Social science (methods)
Gehlbach & Brinkworth (2011)		D, P	D, P		D,P		D	D, P	Social science (psychology)
Guion (1977)	D	D	D	D	D	D	D	D	Social science (psychology)
Hambleton (1984)	D		D,P,T I_{ik}	D,P,T I_{ik}					Education
Haynes et al. (1995)	D	D	D	D	D	D	D		Social science (psychology)
Hemphill & Westie (1950)		D,P	D,P,T I_{ij}	D,P,T I_{ij}			D,P,T r_{xy}		Social science (psychology)
Hinkin & Tracey (1999)			D,P,T ANOVA	D,P,T ANOVA					Social science (methods)
Johnston et al. (2014)			D,P,T T-test	D,P,T DCV	D				Social science (psychology)
Lawshe (1975)			D			D,P,T CVR			Social science (psychology)
MacKenzie et al. (2011)	D		D,P,T ANOVA	D,P,T ANOVA	D				Social science (info systems)
McKenzie et al. (1999)		D	D			D,P,T CVR			Healthcare
Moore & Benbasat (1991)			D,P	D,P,T Kappa		D	D,P,T glb, α	D,P	Social science (info systems)
Petter et al. (2006)			D,P		D				Social science (info systems)

Table 9. Content Validity Recommendations

Polit et al. (2007)			D,P,T CVI, k*						Healthcare (nursing)
Robinson (2018)			D,P,T CVI		D	D	D,P,T α, ω	D,P,T r_{xy}	Social science (management)
Rubio et al. (2003)	D	D,P,T IRA	D,P,T CVI,FVI	P					Social science (social work)
Schriesheim et al (1993)			D,P,T r_{xy}	D,P,T r_{xy}	D				Social science (management)
Sirechi (1998)	D	D	D,P,T I_{ik}	D,P,T I_{ik}	D				Social science (methods)
Straub et al. (2004)			D			D,P,T CVR			Social science (info systems)

D: study describes this aspect of content validity.
P: study provides prescriptive method (process) to evaluate this aspect of content validity.
T: study provides quantitative test

5 Studies for Demonstration and Efficacy

To demonstrate the need for researchers to ensure content validity when developing surveys, we conducted two studies using scales that we developed and adapted from the literature to investigated student behaviors related to course assignments. In the first study, we demonstrate content validity processes that researchers can during instrument development. For this study, we also provide a data-collection and analysis example. In the second study, we conducted an experiment using six survey scales that we cleaned using the content validity methods described in this paper. This second study substantiates efficacy for our recommended pre-study content validity methods.

5.1 Content Validity Pre-study

We engaged pre-study panels in multiple assessment iterations on a survey instrument comprising multiple constructs. We adapted items from existing reflective scales measuring pride-in-craftsmanship and authentic-pride from the literature for use in a future study of student participation in digital assignments. We also included items from the venerable scale for self-efficacy that will not appear in the future study. Appendix C contains the full details of all scales included in the pre-study. These pride and self-efficacy scales are presented in a group for the jurors to match and rate concurrently because they represent orbiting constructs. Table 10 contains an excerpt of the instrument as presented to our pre-study panel.

The population of interest was university-level students for a single course. We sourced domain experts from undergraduate teaching assistants who have taken the target class and who served as laboratory aides. These jurors shared the perspective and vocabulary of the target population as well as an intimate knowledge of the class mechanics. We assembled a second panel with PhD students to serve as academic experts. We instructed these students in survey study methods, and they understood the mechanics of multi-item psychometric surveys. Tables 11, 12, 13, and 14 contain the ratings from the academic jurors. Table 16 contains domain expert ratings from the second round. We used the same mechanics to collect scores and calculate metrics for all panel types.

Table 10. Pilot Study Feedback Form: Round 1

Concept constellation for **pride and self-efficacy**
Pride is a feeling of pleasure and satisfaction from achievement. It is primary emotion that gives self-esteem its affective "kick".

1) **Pride in craftsmanship**: a general work ethic that holds that individuals should produce quality product, regardless of their likes or dislikes, presence (or absence) of supervisors, and should take pride in the results of their efforts. *For this study, this concept establishes a baseline ethic that is broad and persistent across settings.*
2) **Authentic pride**: pride *associated with specific accomplishments* accompanied by genuine feelings of self-worth. *This study attempts to measure pride in class assignments and the work associated with those assignments.*
3) **Self-efficacy**: relates to a people's belief they can successfully implement action and be successful with a specific task. *This study attempts to measure self-efficacy related to class assignments*.
4) **Other** (does not belong here)

Items[±]	**Clarity** 1) Not clear 2) Needs major revision 3) Needs minor revision 4) Clear	**Factor** alignment Choose definition to which item most closely aligns	**Confidence** measures factor 0) No confidence 1) Limited / tenuous 2) Small 3) Moderate 4) Good 5) Very high confidence	**Relevance** to chosen factor 1) Not relevant 2) Somewhat relevant 3) Quite relevant 4) Highly relevant	Comments
A student should do a decent job whether or not his/her teacher is around	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
A student should feel a sense of pride in her/his work	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
There is nothing wrong with doing a poor job on assignments if a person can get away with it	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
Regardless of whether a task is mental or manual, pleasant or unpleasant, it should be performed to the best of one's effort	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
Even if you dislike your assignment, you should do your best	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
Both in class and outside of class, I take pride in the quality of my work	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	

[±] This table includes only items from the PiC scale for illustration purposes. We mixed items from the AP and SE scales with these six PiC items on the form presented to jurors.

Table 11 summarizes juror scores for item clarity, along with the agreement calculation r_{WG}. The second and third items lacked clarity.

Table 11. Item Clarity: Round 1

Item	J1	J2	J3	J4	J5	J6	J7	r_{WG}
DPC1	4	4	4	4	4	4	4	1.00
DPC2	3	3	4	4	1	4	3	***0.09***
DPC3	4	4	4	2	1	4	4	***0.00***
DPC4	4	4	3	4	4	4	4	0.89
DPC5	4	4	4	4	4	4	4	1.00
DPC6	4	3	4	3	4	3	4	0.77

Table 12 summarizes juror scores for item relevance, along with the agreement calculation r_{WG}. The third item lacked internal congruence.

Table 12. Internal Congruence (Relevance): Round 1

Item	J1	J2	J3	J4	J5	J6	J7	r_{WG}
DPC1	4	4	4	2	4	3	4	1.00
DPC2	4		4	4		3	3	0.76
DPC3	4		3	1		3	3	***0.04***
DPC4	4	4	4	4	3	3	4	0.81
DPC5	4	4	4	4	4	4	4	1.00
DPC6	3	3	4	3	2	3	4	0.85

A substantive validity coefficient screening check for external congruence is based on juror's matching items to construct definitions. Table 13 summarizes the results, along with the statistic c_{sv}. The third and fourth items had week alignment, whereas the second item lacked alignment.

Table 13. Factorial Alignment: Round 1

Item	J1	J2	J3	J4	J5	J6	J7	c_{sv}
DPC1	1	1	1	1	1	1	1	1.00
DPC2	1	2	2	2	4	3	1	***-0.43***
DPC3	1	1	2	1	4	4	1	***0.14***
DPC4	1	1	1	3	1	3	1	***0.43***
DPC5	1	1	2	1	1	1	1	0.86
DPC6	1	1	2	2	1	3	1	0.57

We provide a more rigorous test of external congruence based on jurors' confidence that items corresponded to the matched definition. Table 14 summarizes the results, along with the distinctiveness metric htd*. The second and sixth items lacked external congruence, while the third item had moderate congruence.

Table 14. External Congruence: Round 1

Item	J1	J2	J3	J4	J5	J6	J7	htd*
DPC1	5	5	4	4	5	4	4	0.89
DPC2	5	3	4	5		4	3	***-0.23***
DPC3	4	5	2	3		3	3	***0.29***
DPC4	5	5	4	5	4	4	5	0.40
DPC5	5	5	4	5	5	4	5	0.71
DPC6	5	5	4	5	3	4	4	***0.11***

We assessed content consistency for the scale by calculating within-group inter-rater agreement for relevance and confidence. The $r^*_{WG(J)}$ metric across all items was 0.62 for Relevance and 0.81 for confidence. Both measures support content consistency that justifies aggregation.

Many items in this scale had weak or poor content validity. After reviewing the ratings and feedback from the first round with academic jurors, we conducted interviews with domain experts to identify words and phrases that better represented aspects of the latent constructs involved. Table 15 presents the revised feedback instrument. We retained the fourth and fifth items without changes, made minor changes to the first and third items, and substantially revised the second and sixth items.

Table 15. Pilot Study Feedback Form: Round 2

Concept constellation for **pride** and **self-efficacy**
Pride is a feeling of pleasure and satisfaction from achievement. It is primary emotion that gives self-esteem its affective "kick".

1) **Pride in craftsmanship**: does the student share the belief that all students should do quality work regardless of circumstances.
2) **Authentic pride**: does the student have pride in the assignments (the work, the file, and answers) they submit in class.
3) **Self-efficacy**: does the student have the belief they can successfully do class assignments.
4) **Other** (does not belong here)

Items	**Clarity** 1) Not clear 2) Needs major revision 3) Needs minor revision 4) Clear	**Factor** alignment Choose definition to which item most closely aligns	**Confidence** measures factor 1) No confidence 2) Limited / tenuous 3) Small 4) Moderate 5) Good 6) Very high confidence	**Relevance** to chosen factor 1) Not relevant 2) Somewhat relevant 3) Quite relevant 4) Highly relevant	Comments
All students should do a decent job even when ***the*** teacher is not ***present.***	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
Everyone should have pride in the quality of ***their*** work in ***all situations.***	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
There is nothing wrong with doing a poor job ***if*** a person can get away with it.	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
Regardless of whether a task is mental or manual, pleasant or unpleasant, it should be performed to the best of one's effort.	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
Even if you dislike your assignment, you should do your best.	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	
I take ***pride in the work I do outside of school.***	1 2 3 4	1 2 3 4	0 1 2 3 4 5	1 2 3 4	

We reassembled the panel of domain experts approximately six weeks after round one to gather a second round of feedback. Table 16 presents content validity metrics for clarity, internal congruence, external congruence, and consistency. Guided by feedback from the panel, we dropped the second and third items and recalculated content consistency metrics.

Table 16. Round 2 Statistics

Item	Clarity	Internal congruence	External congruence	External congruence	Content consistency (relevance)	Content consistency (confidence)
	r_{WG}	r_{WG}	c_{sv}	htd*	$r^*_{WG(J)}$	$r^*_{WG(J)}$
CPC1	1.00	0.84	1.00	0.88	0.68 [0.96]±	0.61 [0.94]±
CPC2	***0.76***	***0.00***	***-0.60***	***-0.36***		
CPC3	1.00	**0.60**	***0.20***	**0.28**		
CPC4	1.00	0.84	0.60	0.52		
CPC5	1.00	0.84	1.00	0.88		
CPC6	1.00	1.00	0.60	0.52		

± Consistency ratings in [brackets] constitute recalculated values after we dropped the second and third items

5.2 Field Experiment

To demonstrate the efficacy of the content validity methods that we used in the pre-study, we conducted a randomized field experiment that involved six reflective survey scales. Three scales (see Tables C1, C2, and C3 in Appendix C) involved neighboring constructs: self-efficacy, pride-in-craftsmanship, and authentic-pride. We developed these scales using conventional practices that researchers follow to adapt existing scales in published studies. During adaptation, we adjusted each question's context to the target study context and adjusted the referents so they pertained to the intended subjects. We reviewed and adjusted questions prior to engaging pre-study jurors or final-study informants.

We devised new scales that represented three dimensions (controlling-a-target, knowing-a-target, and self-identifying-with-a-target) of a theorized second-order construct, psychological-ownership (see Tables C4, C5, and C6 in Appendix C). We devised these scale items using conventional practices that researchers follow in reviewing the literature and aligning questions to conceptual ideas linked to the theory. We reviewed and adjusted questions prior to engaging pre-study jurors or final-study informants.

5.2.1 Method and Procedures

To develop the survey items, we followed the following multi-step process:

1) Initial literature search: we searched the literature to identify the theoretical basis for constructs of interest, identify and adapt existing scales where available (self-efficacy, pride-in-craftsmanship, and authentic-pride), and develop new scales where suitable scales did not exist (controlling-a-target, knowing-a-target, and self-identifying-with-a-target). When developing new scales, we collected keywords guided by relevant theory and the literature and created a pool of candidate questions.

2) Qualitative review: we assembled a panel of undergraduate teaching assistants (n = 10) to serve as domain experts. These jurors represent a sample of the target population for a final study with first-hand knowledge of the assignments and instructional methods employed in the target class. This panel reviewed the survey instrument and questions for clarity, completeness, relevance and adequacy. Jurors provided categorical feedback (yes/no) and extensive free-form qualitative feedback to guide rewording of items to the target context.

3) First revision: using feedback from the qualitative review, scales were revised by the study investigators. The resulting scales appear in Appendix C with the designation "dirty".

4) Content validity pre-study (first round): a panel of PhD students (n = 7) was trained on content validity and pre-study methods and serve as academic experts. This panel performed a pen-and-paper rating of constructs and items. The instrument was analyzed for clarity, congruence, and dependability.

5) Interview with domain experts (n = 5): domain experts were asked to review the construct definitions and discuss problem areas for specific questions. Jurors made specific recommendations for wording changes that they believed better align items with the target constructs.

6) Second revision: using feedback from pre-study first round and Interviews with domain experts, the scales were again revised by the study investigators.

7) Content validity pre-study (second round): the panel of undergraduate teaching assistants (n = 7) was reassembled to serve as domain experts. This panel performed an electronic rating of constructs and items. The instrument was analyzed for clarity, congruence, and dependability.

8) Third revision: we made final changes to survey items. We made no additional wording changes but dropped some low-performing items. The resulting scales appear in Appendix C with the designation "clean".

To demonstrate the proposed content validity assessment exercise's efficacy, we performed a randomized field experiment comparing the two survey instruments. We invited 920 students who took a single class during one semester to participate in the experiment as an extra credit exercise. We also offered the students an alternate extra credit task of equal value. We administered the survey through Qualtrics as an A/B experiment where we randomly assigned students to either the "dirty" survey scales or the "clean" survey scales. In total, 725 students agreed to participate in the study (i.e., provided informed consent). We discarded eight incomplete responses. We also discarded 221 responses that failed either or both of two attention check questions we included in the survey. The final usable sample for the clean survey instrument comprised 245 responses, and the final usable sample for the dirty survey instrument comprised 251 responses.

5.2.2 Results

We first examined the dirty scales to demonstrate construct validity using both PLS (WarpPLS 6.0) and CFA (R 3.6.1, lavaan 0.6-5, semTools 0.5-2). Table 17 details the key statistics. Many items failed the tests for indicator reliability since they loaded below the common threshold of 0.70. In many cases items loaded below the exploratory threshold of 0.40 (Hair et al. 2012) or lacked statistical significance. Three constructs failed the test for internal consistency with a composite reliability (also known as Dillon-Goldstein's rho, omega ω_u) score below the common threshold of 0.70 and, in some cases, below the exploratory research limit of 0.60 (Hair et al., 2012). Four constructs failed the test for convergent validity with an AVE below the threshold of 0.50 (Hair et al., 2012). The CFA statistics also revealed that many items failed the test for indicator reliability. In addition, two constructs failed the test for convergent validity, and the goodness of fit indices were marginal.

At this point, many researchers begin removing items as they seek to find a measurement model with acceptable construct validity in order to perform a path analysis. Typically, they do so by iteratively omitting outlier items or items with particularly poor indicator reliability (loading). We made iterative deletions until the measurement models demonstrated adequate construct validity to support model (path and hypothesis) testing. Table 18 presents the measurement model statistics for this revised scale designated "reduced dirty". The analysis below exposes evidence of poor content validity, which suggests that path-model testing would include substantial systemic measurement error.

Table 19 reports similar PLS and CFA measurement model statistics for the "clean" constructs. Items dropped in step eight of the pre-study are designated with a strikethrough. All scales demonstrated good convergent validity across all metrics.

Table 17. Construct Validity for "Dirty" Constructs

PLS					CFA				
Items	Loading†	pVal	ω_u (α)	AVE	Loading†	pVal	ω_h (α)	AVE	GoF
DSE_1	0.860	<0.001	0.866 (0.80)	0.69	0.800		0.811 (0.80)	0.58	SRMR = 0.097 CFI = **0.834** RMSEA = **0.120**
DSE_2	0.840	<0.001			0.868	<0.00			
DSE_3	0.709	0.011			0.635	<0.00			
DPC_1	*0.641*	**0.187**	**0.523** (0.71)	**0.17**	0.742		0.747 (0.64)	**0.46**	
DPC_2	*0.364*	**0.360**			0.778	<0.00			
DPC_3	0.907	**0.228**			**-0.369**	<0.00			
DPC_4	*0.660*	**0.163**			0.724	<0.00			
DPC_5	*0.577*	**0.229**			0.796	<0.00			
DPC_6	*0.477*	**0.271**			0.757	<0.00			
DAP_1	0.707	**0.092**	0.863 (0.93)	**0.38**	0.844		0.885 (0.93)	0.55	
DAP_2	0.779	**0.052**			0.944	<0.00			
DAP_3	0.713	**0.066**			0.938	<0.00			
DAP_4	*0.666*	**0.086**			0.803	<0.00			
DAP_5	0.749	**0.045**			0.867	<0.00			
DAP_6	0.729	0.023			*0.596*	<0.00			
DAP_7	0.759	**0.052**			0.782	<0.00			
DAP_8	0.761	**0.039**			0.860	<0.00			
DAP_9	0.844	0.016			*0.575*	<0.00			
DAP_10	0.889	0.045			**0.374**	<0.00			
DAP_11	0.963	**0.054**			**0.357**	<0.00			
DOC_1	0.*523*	0.005	0.925 (0.90)	0.65	**0.298**		0.919 (0.91)	0.66	
DOC_2	0.912	<0.001			0.799	<0.00			
DOC_3	0.922	<0.001			0.836	<0.00			
DOC_4	0.932	<0.001			0.940	<0.00			
DOC_5	0.928	<0.001			0.943	<0.00			
DOC_6	0.913	<0.001			0.930	<0.00			
DOC_7	0.827	<0.001			0.542	<0.00			
DOK_1	0.756	0.020	**0.066** (0.82)	**0.15**	0.703		0.828 (0.82)	0.55	SRMR = 0.091 CFI = **0.860** RMSEA = 0.098
DOK_2	**0.323**	**0.368**			0.884	<0.00			
DOK_3	**-0.232**	**0.429**			0.660	<0.00			
DOK_4	*-0.404*	**0.393**			0.680	<0.00			
DOSI_1	0.932	**0.175**	**0.267** (0.77)	**0.19**	**0.222**		0.771 (0.72)	**0.36**	
DOSI_2	0.887	**0.124**			*0.548*	<0.00			
DOSI_3	0.892	**0.129**			*0.644*	<0.00			
DOSI_4	0.645	**0.110**			*0.569*	<0.00			
DOSI_5	-0.968	**0.219**			**0.022**	**0.747**			
DOSI_6	**-0.135**	**0.426**			0.761	<0.00			
DOSI_7	**-0.046**	**0.479**			0.834	<0.00			
DOSI_8	**0.396**	**0.282**			0.790	<0.00			
DOSI_9	**-0.355**	**0.388**			**0.335**	<0.00			

† standardized loadings. ω_u is a unidimensional version of omega mathematically equivalent to Dillon-Goldstein rho composite reliability. α is Cronbach's coefficient alpha (assumes tau-equivalence). ω_h is McDonald's coefficient omega. Italics indicate marginally acceptable. Bold indicates not acceptable.

Table 18. Construct Validity for "Reduced Dirty" Constructs

PLS					CFA				
Items	Loading†	pVal	ωu (α)	AVE	Loading†	pVal	ωh (α)	AVE	GoF
DSE_1	0.779	<0.001	0.866 (0.80)	0.69	0.800		0.811 (0.80)	0.58	SRMR = 0.078 CFI = 0.910 RMSEA = 0.099
DSE_2	0.753	<0.001			0.868	<0.001			
DSE_3	*0.573*	0.010			*0.633*	<0.001			
DPC_1	0.775	0.009	0.876 (0.89)	0.56	0.750		0.871 (0.87)	0.57	
DPC_2	*0.622*	0.045			0.785	<0.001			
~~DPC_3~~									
DPC_4	0.799	0.009			0.719	<0.001			
DPC_5	0.774	0.003			0.790	<0.001			
DPC_6	0.720	0.004			0.755	<0.001			
DAP_1	*0.589*	0.025	0.878 (0.94)	*0.45*	0.846		0.931 (0.94)	0.65	
DAP_2	*0.648*	0.007			0.946	<0.001			
DAP_3	*0.616*	0.010			0.941	<0.001			
DAP_4	*0.533*	0.036			0.804	<0.001			
DAP_5	*0.652*	0.005			0.867	<0.001			
DAP_6	*0.667*	0.003			*0.592*	<0.001			
DAP_7	*0.600*	0.007			0.781	<0.001			
DAP_8	*0.616*	0.005			0.858	<0.001			
DAP_9	0.784	0.024			*0.562*	<0.001			
~~DAP_10~~									
~~DAP_11~~									
~~DOC_1~~			0.945 (0.93)	0.74			0.926 (0.93)	0.69	SRMR =0.048 CFI =0.966 RMSEA = 0.062
DOC_2	0.880	<0.001			0.796				
DOC_3	0.880	<0.001			0.834	<0.001			
DOC_4	0.898	<0.001			0.940	<0.001			
DOC_5	0.899	<0.001			0.945	<0.001			
DOC_6	0.999	<0.001			0.930	<0.001			
DOC_7	0.839	<0.001			*0.541*	<0.001			
~~DOK_1~~			0.818 (0.78)	0.60			0.786 (0.78)	0.55	
DOK_2	*0.626*	0.023			0.829				
DOK_3	0.719	0.013			*0.664*	<0.001			
DOK_4	0.763	0.017			0.720	<0.001			
DOSI_1	0.877	0.003	0.829 (0.78)	0.50	*0.436*		0.779 (0.73)	*0.43*	
DOSI_2	0.794	<0.001			0.769	<0.001			
DOSI_3	0.813	<0.001			0.792	<0.001			
DOSI_4	*0.614*	0.022			0.703	<0.001			
~~DOSI_5~~									
~~DOSI_6~~									
~~DOSI_7~~									
DOSI_8	*0.636*	0.020			*0.552*	<0.001			
~~DOSI_9~~									

We removed items with a strikethrough after data collection to improve construct validity. † Standardized loadings. ω_u is a unidimensional version of omega, mathematically equivalent to Dillon-Goldstein rho composite reliability . α is Cronbach's coefficient alpha (assumes tau-equivalence). ω_h is McDonald's coefficient omega. Italics indicate marginally acceptable.

Table 19. Construct Validity for "Clean" Constructs

PLS					CFA				
Items	Loading†	pVal	ω_u (α)	AVE	Loading†	pVal	ωh (α)	AVE	GoF
CSE_1	0.784	0.003	0.882 (0.82)	0.72	0.780		0.820 (0.82)	0.61	
CSE_2	*0.653*	0.004			0.888	<0.001			
CSE_3	0.748	0.006			0.660	<0.001			
CPC_1	0.802	0.002	0.856 (0.79)	0.60	0.586		0.787 (0.79)	0.5	SRMR =0.045
~~CPC_2~~									
~~CPC_3~~									
CPC_4	0.720	<0.001			0.801	<0.001			CFI =0.970
CPC_5	0.755	<0.001			0.800	<0.001			
CPC_6	0.761	<0.001			0.603	<0.001			RMSEA =0.068
CAP_1	0.767	<0.001	0.964 (0.96)	0.81	0.794		0.957 (0.96)	0.79	
CAP_2	0.734	<0.001			0.924	<0.001			
CAP_3	0.720	<0.001			0.945	<0.001			
CAP_4	0.713	<0.001			0.899	<0.001			
CAP_5	*0.697*	<0.001			0.852	<0.001			
CAP_6	0.745	<0.001			0.899	<0.001			
~~CAP_7~~									
~~COC_1~~			0.955 (0.95)	0.81			0.937 (0.95)	0.78	
COC_2	0.966	<0.001			0.777				
COC_3	0.964	<0.001			0.837	<0.001			
COC_4	0.941	<0.001			0.879	<0.001			
COC_5	0.927	<0.001			0.952	<0.001			
COC_6	0.918	<0.001			0.934	<0.001			
COK_1	*0.676*	0.002	0.886 (0.86)	0.66	0.787		0.856 (0.85)	0.60	SRMR =0.052
COK_2	0.739	0.004			0.782	<0.001			CFI =0.935
COK_3	0.762	0.002			0.717	<0.001			
COK_4	0.685	0.005			0.821	<0.001			RMSEA =0.095
~~COSI_1~~			0.903 (0.90)	0.65			0.898 (0.89)	0.64	
COSI_2	0.657	0.007			0.732				
~~COSI_3~~									
COSI_4	0.649	0.008			0.797	<0.001			
COSI_5	0.737	0.002			0.889	<0.001			
COSI_6	0.751	0.006			0.756	<0.001			
COSI_7	0.725	0.005			0.800	<0.001			

We removed items with strikethrough from the survey scale in the eighth step in the development process. † Standardized loadings. ω_u is a unidimensional version of omega mathematically equivalent to Dillon-Goldstein rho composite reliability . α is Cronbach's coefficient alpha (assumes tau-equivalence). ω_h is McDonald's coefficient omega . Italics indicate marginally acceptable.

These construct validity statistics provide strong support for the contention that the pre-study content validity methods that we present in this paper represent a genuine improvement. To visualize the differences, we performed a principle component analysis for each construct cluster using Stata statistical software. Figure 6 presents the principle component loading plots calculated using maximum likelihood factor analysis and oblique oblimin rotation.

Figure 6. Loading Plots

The plot of dirty items visually reveals that items from both the pride in craftsmanship and the authentic pride scales correlated more with the self-efficacy items than with their intended construct. In addition, the core self-efficacy and authentic pride constructs seem to have merged into a larger ambiguous construct. The reduced dirty scales omitted items that we removed after data collection to improve convergent validity. These plots of reduced dirty items show improvement, although some items for authentic pride seem to have joined the self-efficacy construct. In contrast, the clean items congregated in three tidy groups and, thus, depict good congruence.

The three scales in the psychological ownership cluster tell a similar story. The plot of dirty items visually reveals that items from all three scales mingled with a possible core for ownership control. In contrast, the clean items congregated in three tidy groups and, thus, depict good congruence.

We supplement the visually compelling loading plots with some more objective tests to calculate the number of distinct factors. Each construct cluster involved three orbiting constructs. Either too few or too many factors suggest that the scale lacks content validity. We employed three methods to calculate the number of distinct factors: the parallel analysis method (Horn, 1965), the minimum average partial (MAP) method (Velicer, 1976), and the Bayesian information criteria method. Researchers have demonstrated each method in simulation studies to be highly accurate when analyzing data with an item-to-factor ratio consistent with that involved in our experiment (Lorenzo-Seva, Timmerman, & Kiers, 2011; Zwick & Velicer, 1986). Table 20 reports findings from the distinct factor analyses. Only the clean item scales resulted in models with three factors, which suggests poor content validity for the dirty and reduced dirty scales.

Table 20. Empirical Estimation of Number of Constructs

Scale	Number of constructs		
	Horn's Parallel Analysis†	Velicer's MAP‡	BIC*
Pride and self-efficacy custer (dirty)	6	4	6
Pride and self-efficacy cluster (reduced dirty)	4	4	5
Pride and self-efficacy cluster (clean)	**3**	**3**	**3**
Psychological ownership cluster (dirty)	5	4	5
Psychological ownership cluster (reduced dirty)	4	4	4
Psychological ownership cluster (clean)	**3**	**3**	**3**

† We performed Horn's parallel analysis with CFA estimation (R package paran 1.5.2)
‡ We calculated Velicer's MAP using maximum likelihood FA factoring method with oblimin rotation. We calculated correlations with robust Spearman's rank-based measure of association (R package psych 1.8.12)
* We calculated Bayesian information criterion (BIC) calculated using maximum likelihood FA factoring method with oblimin rotation. We calculated correlations with robust Spearman's rank-based measure of association (R package psych 1.8.12)

Tests to determine the number of constructs expose contamination when the number does not match the theorized model. However, even when the number of constructs matches the theorized model, items may contain unique influence from many constructs. The Schmid-Leiman solution (Schmid & Leiman, 1957) generates a multidimensional factor model that parcels out common factor bias and exposes the cross-factor contamination (Wolff & Preising, 2005). Figure 7 depicts Schmid-Leiman factor solutions for the scales assessed in this experiment. We performed this analysis using R (psych 1.8.12) for three theorized factors using the principle component method to reveal each factor's unique influence on all items. The dirty reduced scales included significant contamination across factors and items that better represent an incorrect factor.

Both subjective and objective assessments of the data that we collected in this experiment strongly support the value of assessing scales for content validity in a pre-study exercise. Scales may contain contamination and other forms of systemic measurement error even when conventional construct validity metrics allow use.

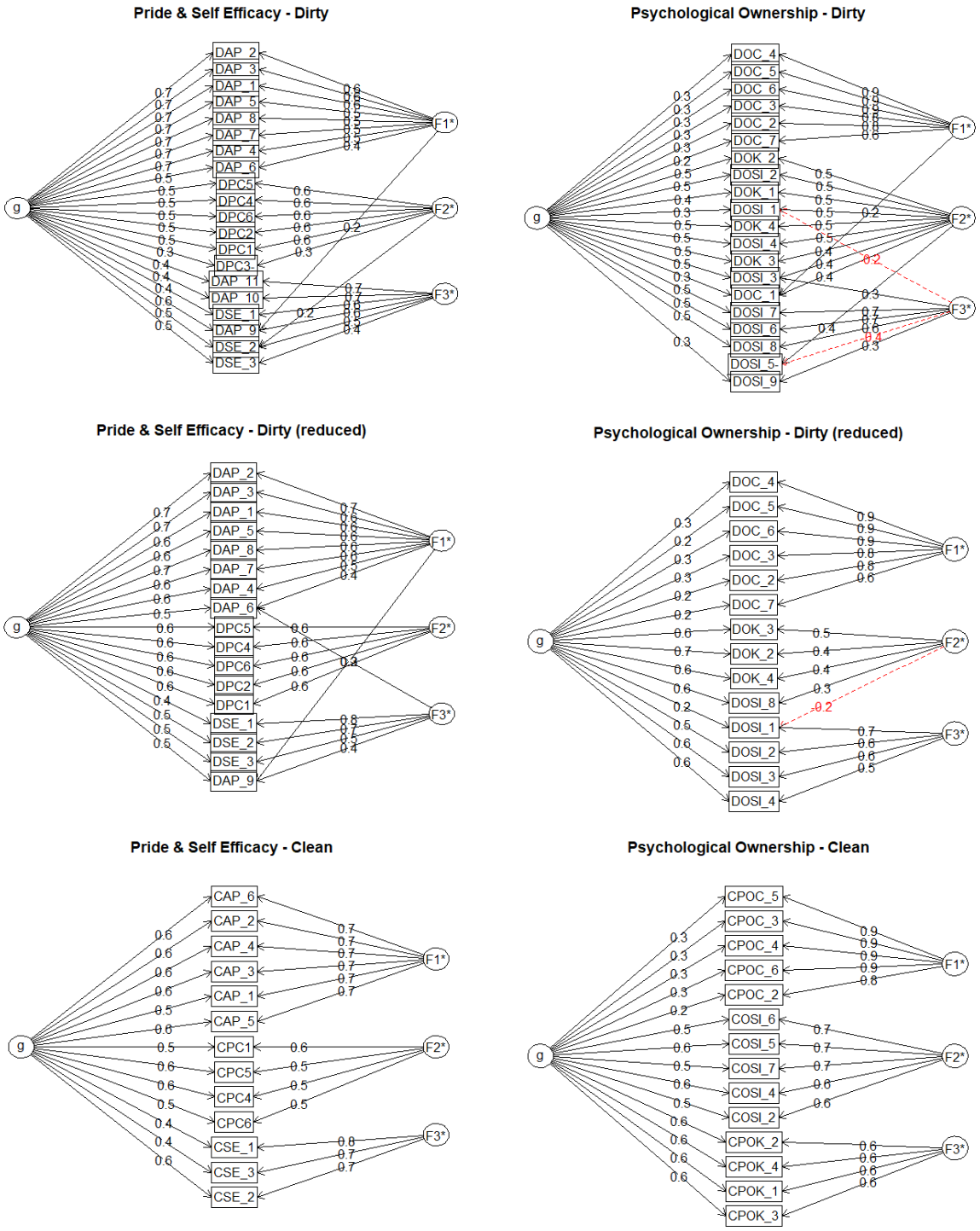

Figure 7. Schmid-Leiman Solution Schematic Diagrams

6 Discussion

This paper identifies content validity considerations for survey scales. Even though researchers often overlook content validity, this study presents efficacy evidence demonstrating its importance for survey studies in the information systems field. Content validity is a higher standard than face validity, which requires only one observer to subjectively conclude that an item bears a common-sense relationship to a construct. Long before consensus agreement methods emerged to evaluate content validity, Mosser

(1947) dismissed face validity as a "pernicious fallacy" (p. 208) and recommended that researchers "banish" the term "to outer darkness" (p. 191).

We decompose the landscape of content validity into its constituent facets, with precise definitions to unite the discordant labels that appear in literature. In curating methods suitable to evaluate content validity, we detail papers that discuss aspects of content validity, that describe evaluation processes, and that describe quantitative tests. We match each facet with best practices for examining content validity along with suitable quantitative metrics. This paper also presents strengths and weaknesses of competing methods to guide researchers who seek to demonstrate content validity.

Existing content validity methods face challenges in at least two areas. The first challenge is implementing fully crossed pre-study designs when rating congruence of survey items. We present a nested pre-study design and propose a new quantitative metric, htd*, that researchers can use to simultaneously evaluate both internal congruence and external congruence. The second challenge is establishing adequacy for formative constructs. We recommend a qualitative approach for indicator sufficiency that involves an iterative nominal group technique. We further recommend that researchers transform indicator parsimony into a hierarchical exercise in which they retain indicators with the highest congruence ratings for individual dimensions.

In reviewing recent IS literature that has employed survey methods, we developed a snapshot of the state of practice for assessing content validity. Although IS researchers appear interested in achieving content validity, they have tended to employ inconsistent and suboptimal methods. To help researchers evaluate content validity, we demonstrate suitable pre-study rating methods and quantitative metrics to evaluate reflective scales. A suitable way to reduce subjectivity concerns is to employ pre-study assessments with appropriate metrics for each facet of content validity (Stemler, 2004). An iterative "rinse and repeat" approach builds confidence that the resulting instrument is stable. We also conducted a field experiment to demonstrate the improvement achieved using these methods.

Our experiment—particularly the portion involving adapted scales in the pride and self-efficacy cluster—implies that researchers should revalidate adapted scales. This is not a new revelation. MacKenzie et al. (2011, p. 296) note that their 10-step procedure for scale development applies when a scholar does "construct conceptualization (or reconceptualization of an existing construct)." Haynes et al. (1995, p. 241) take a particularly blunt stance on this subject in noting:

> *Assessment instruments can have different functions, and indices of validity for one function of an instrument are not necessarily generalizable to other functions of the instrument. Consequently, validity indices are conditional—they pertain to an assessment instrument, when used for a particular purpose.*

Furthermore, they add that "content validity often degrades over time as new data are acquired and theories about the target construct evolve".

In summary, in this paper, we:

1) Present a tutorial that describes the mechanisms of multi-item psychometric scales to illustrate content validity considerations.
2) Precisely define each facet of content validity.
3) Summarize the content validity literature and identify papers that discuss facets of content validity, describe evaluation processes, and describe quantitative tests.
4) Present strengths and weaknesses of evaluation processes applicable to each facet of content validity and make specific recommendations based on them.
5) Identify strengths and weaknesses of naïve jurists, academic experts and domain experts for pre-study panels and make specific recommendations based on them.
6) Present suitable quantitative metrics that align with each content validity evaluation process.
7) Describe a new nested design for pre-study congruence evaluation that addresses challenges with the fully crossed design.
8) Describe a new quantitative metric to calculate congruence when using nested designs.

9) Propose a qualitative nominal group technique for establishing indicator sufficiency for formative scales.

10) Propose using the htd* metric to select the most congruent item when pruning formative scales to ensure indicator parsimony.

11) Demonstrate, via an example, rating-based methods for pre-study survey assessment that employ the quantitative metrics we recommend.

12) Demonstrate the efficacy of the methods we propose through a field experiment.

Taken together, this paper broadly examines content validity for survey scales.

6.1 Formative Indices, Sometimes Just Say "No"

The adequacy facet of content validity establishes a high standard for formative indices. Indicator sufficiency does not lend itself to quantitative measures for conventional subject matter expert panels or ad hoc academic panels. Identifying a complete list of a latent construct's important domains is often the intense focus of multi-year research programs. For example, DeLone and McLean (2016) recount a journey in which they reviewed literature published over a ten-year span to establish an initial taxonomy of IS success factors. They refined the taxonomy with a hierarchical view of influence that resulted in six dimensions: system quality, information quality, use, user satisfaction, individual impact, and organization impact. Shortly after publication (DeLone & McLean, 1992), the IS community engaged in discussion and proposed numerous modifications. For ten years, a debate raged as numerous studies and opinion papers analyzed the topic in various IS journals. The debate culminated in a revised IS success model (DeLone & McLean, 2003) that added two new dimensions: service quality and intention to use. The revised model also collapsed the individual impact and organization impact dimensions into a single dimension called net benefits. Thus, establishing indicator sufficiency for IS success involved dozens of studies over twenty years. Similar journeys exist (or are still underway) for other formative constructs. This is even true for the Big Five personality traits. Early research started in the 1880s with agreement on a five-factor model finally appearing in 1980, almost 100 years later.

The speed with which information and communication technologies change complicates efforts to sufficiently identify IS-specific constructs. For example, Marakas, Johnson, and Clay (2007) argue that researchers should measure computer self-efficacy (CSE) as both a technology- and context-specific formative construct. The dynamic nature of Information and computer technologies is particularly relevant for a construct such as CSE. Consider the specific area of "telephone self-efficacy". The facets of this construct differ today compared to the 1960s when rotary phones represented the norm. Each technology advance, from pulse dial to touch-tone to mobile to multi-function mobile and smart phones, necessitates new dimensions to operationalize a formative index for telephone self-efficacy. Given the speed with which the entire information and communication technology landscape changes, a 20- or 100-year project to establish the dimensions of a new formative construct in the IS domain amounts to a Sisyphean task.

This brings us to the ongoing debate about the suitability of formative constructs. Petter et al. (2007) note that noumena are neither inherently formative nor reflective. Some scholars suggest using multiple-indicator, multiple-cause (MMIC) models (Hauser & Goldberger, 1971; Cenfetelli & Bassillier, 2009) to simultaneously collect reflective items and formative indicators for a single latent construct. Researchers have applied this approach to IS constructs such as IS success (e.g., Gable, Sedera, & Chan, 2008). Recognizing that one can measure a latent construct with both reflective scales and formative indexes leads directly into a debate where some scholars suggest setting aside formative indexes in favor of reflective scales (Bagozzi, 2011; Edwards, 2011; Hardin, Chhang, & Fuller, 2008). It is beyond the scope of this paper to resolve this debate, which often hinges on the objectives and context of an individual study, but we do add an additional consideration to the discussion. A single study cannot easily accomplish the journey to establish indicator sufficiency content validity. Researchers who attempt to measure a latent construct using formative measures must often rely on decades of prior work to gain any confidence in a claim of indicator sufficiency. Even this foundation is unstable due to the dynamic nature and pace of change among information and communication technologies.

Indicator sufficiency exposes another problem for formative constructs: structural equation modeling places the error term at the construct level for formative latent factors such that it "represents the impact of all remaining causes other than those represented by the indicators included in the model" (Diamantopoulos, 2006, p. 11). Beyond formalizing the need to measure all facets of a latent construct,

this feature of SEM means that existing analysis techniques do not have the capacity to simultaneously account for other forms of measurement error associated with formative constructs. SEM does not model other forms of measurement error but rather assumes they are absent, which is wildly optimistic when considering issues such as item clarity. As a result, other forms of measurement error make a construct's meaning progressively ambiguous (Diamantopoulos, Riefler, & Roth, 2008).

6.2 Limitations

Providing evidence for dependability is a multifaceted undertaking. The methods that we present here assume a homogenous target population. We do not explicitly address the challenge of standard presentation and interpretation for heterogeneous populations. Standard presentation poses unique concern when informants span different cultural backgrounds and primary languages where individual words convey different meaning. Assuring standard presentation and interpretation grows in complexity when considering accessibility for communication handicaps such as blind or deaf informants. Researchers must apply care to assure their survey instruments present content with equivalent meaning for all target subgroups. One option is to include individuals who represent each subgroup in pre-study panels. Depending on the gap between subgroups, researchers may need to conduct translation and back-translation exercises during an iterative pre-study to achieve content consistency.

The demonstration and experiment in this paper involve processes and metrics suitable for reflective survey scales. The demonstration and experiment do not include formative construct methods. Formative scales require qualitative methods, such as structured interviews and the nominal group technique, to establish indicator sufficiency. However, this paper focuses on advancing quantitative methods.

An additional limitation involves the htd* metric that we propose for assessing congruence. We "borrowed" the evaluation criteria from the simulation exercise that Colquitt et al. (2019) performed. This simulation did not explicitly model the nested design we propose for htd*. As a result, researchers should use the proposed evaluation criteria with this understanding. Furthermore, researchers should not make the decision to include, modify, or exclude individual items and indicators exclusively based on these metrics. The relative ratings guide the instrument-development process and collectively build confidence in content validity claims. Crafting efficient survey items remains an art that draws on the talents of a researcher to adapt the vocabulary of their target population to the noumenon and nomological network of their study. We recommend processes and statistics to guide researchers and provide confidence to study reviewers and consumers. Positivist IS studies are constrained by the stochastic nature of the various techniques employed. Using repeatable methods, such as the ones we recommend here for content validity, can reduce measurement error to make findings more robust. However, researchers must always interpret them within the limits of precision inherent in choices such as $\alpha \leq 0.05$.

6.3 Future Research

Although we propose a variant of the htd metric for evaluating nested designs of external validity scores, a simulation study to establish evaluation criteria specifically for htd* would improve the confidence of researchers during the pre-study process. Such a simulation could bolster confidence when htd* is applied to both congruence and item parsimony.

The a_{wg} index represents another new metric that may apply to several content validity facets (Brown & Hauenstein, 2005). This metric originated in multilevel research studies, but might suit test-retest assessments that compare agreement across two times and places. This metric might also have applicability to analyses that aggregate scores for evaluating internal consistency at the scale level.

This paper presents assessment methods developed in specific knowledge domains, such as healthcare or human resources. The healthcare domain is pursuing additional new metrics with possible application to content validity. For example, the coefficient of repeatability (CR) (Vaz, Falkmer, Passmore, Parsons, & Andreou, 2013), measures "absolute reliability". We present the Pearson's correlation coefficient (r) and interclass correlation coefficient (ICC), which fall into the "relative reliability" category. Thus, future researchers could explore the extent to which absolute reliability metrics apply to content validity.

The psychometric multi-item scales that we discuss in this paper operationalize measures adhering to classical test theory (CTT). CTT scales measure the level of a latent trait. Researchers in the behavioral sciences use CTT methods to evaluate theories that characterize groups and populations. Item response theory (IRT) offers a different perspective on multi-item scales and focus on assessing item-difficulty and person-trait parameters (Osteen, 2010). IRT leverages adaptive computer-driven surveys to characterize

an individual in a population. In the education domain, IRT surveys allow one to determine specific competency levels (of mathematics or a foreign language) to assign a grade or place a student into a curriculum at an appropriate level. In the healthcare domain, IRT allows a clinician to determine a specific individual's state to guide the selection of a customized treatment plan. An IRT scale uses multiple questions to adaptively triangulate the subjects' unique level. IRT scales require significantly larger pools of questions where success or failure on a given question establishes a floor or ceiling for the person-trait and guides selection of the next question to eventually pinpoint the person-trait level. Researchers require much larger jury pools and pilot study cohorts to establish the characteristics of an individual question and the collective efficacy of an entire question pool. The IRT scale-development process is vulnerable to content validity issues in the same way as CTT. Establishing content validity using pre-study panels allows researchers to minimize the size of pilot study cohorts and the number of pilot studies necessary to calibrate the difficulty of individual items and person-trait profiles for an adaptive item pool. Future research needs to address the unique content validity characteristics that apply to establishing difficulty for individual items.

6.4 Recommendations for Authors and Reviewers

In reviewing research that *MIS Quarterly*, *Information Systems Research*, the *Journal of Management Information Systems*, and the *Journal of the Association for Information Systems* published in 2018 and 2019, we found that the IS community has often used fragmented and incomplete content validity practices. By describing the full landscape of content validity and exposing empirical methods to quantify each facet of content validity, we hope to encourage more researchers who choose to address content validity to transition to objectively reproducible methods using quantitative metrics.

During the transition, many scholars and reviewers may find themselves attempting to navigate the peer-review process for a study that lacks pre-study efforts for content validity. We suggest a few exemplars as a guide for handling this situation. Consider the assumption of normally distributed data, which underlies many statistical techniques. While many studies document normality tests, others use analysis methods purported to be robust even in the presence of non-normally distributed data. Still others argue that the assumption of normality is less a concern as sample size grows. Some published studies simply remain silent on the subject and neglect any attempt to demonstrate that normality assumptions are met. Authors and reviewers currently make case-by-case judgement calls to set the bar high or low for an individual study. Content validity may warrant a similar approach, whereby judgement calls are guided by the importance and intended application of a study's conclusion. Hambleton (1984, p. 205) observes that "accumulating validation evidence is a never-ending process. The amount of time and energy that is expended in the direction of validation of test scores must be consistent with the importance of the testing program".

A second exemplar relates to conventional guidelines for construct validity. Exploratory studies may sometimes defend a lower threshold of validity than theory confirmation studies (Hair et al., 2012). Studies that have greater consequence justify that authors rigorously address content validity to build confidence that measures do in fact capture the noumena that they purport to capture.

A third exemplar involves the handling of internal validity and the difference between causation, correlation, and association (Altman & Krzywinski, 2015). While the scientific community values proof of causation to be far more meaningful than proof of correlation, the community has not relegated correlation findings to the dustbin of history. Instead, the scientific community embraces studies that identify correlations with the caveat that they temper claims by explicitly acknowledging methodological limitations. Furthermore, authors are reminded to limit inferences that involve internal validity (causation).

From these exemplars, we conclude that studies that provide limited evidence in the content domain may still represent useful knowledge that is worthy of documentation and dissemination. However, authors should transparently address their methodological limitations. We offer several observations to help scholars with content validity:

1) Multi-item psychometric scales involve item samples that researchers take from a population of possible questions that could measure aspects of a noumenon. As Cronbach and Meehl (1955, p. 281) observe: "content validity is ordinarily established deductively by defining a universe of items and sampling systematically within this universe to establish the test". Without evidence that the systematic sampling has generated a valid sample of the noumenon,

researchers should use care when making inferences from the scale (the *implied* noumenon) to the theoretical construct (the *real* noumenon).

2) Content-related evidence of validity is grounded in judgement but is not the same thing as "face validity". Face validity requires only the subjective evaluation from any casual observer that an item bears a common-sense relationship to a construct. Establishing content validity requires one to systematically evaluate many facets of a measurement instrument. Quantifiable methods to evaluate "fit for purpose" by juror consensus represent a best practice to build confidence in a claim that scales measure their target noumena.

3) Evidence from one facet of content validity (e.g., congruence) does not serve as evidence for others (e.g., adequacy). Claims that inferences from data collected using a scale generalize to the theoretical noumenon broadly gain credibility from aggregated evidence across all facets. Authors should proceed cautiously when making claims and inferences based on evidence from a single facet.

4) Researchers should exercise caution when reusing scales from published studies. Just because a scale appears in a prior study does not necessarily signal validity. Researchers should be mindful that many published studies do not address content validity in a complete or appropriate manner as we illustrate in Table A2. Furthermore, when researchers validate scales, they do so in a specific setting. Validation studies should make clear statements about the bounds in which validity claims hold. Researchers who move a scale outside these bounds should make a case for their degree of confidence in claims and inferences applicable to the new setting (Seddon & Scheepers, 2012). A new setting could include a change in the studied population, a change in the environment (contextual or temporal), a change in the nomological network, or even a change in the formal definition of constructs under consideration. The change in setting often warrants new pre-studies to validate an alternate vocabulary applicable to the target population, or even a different number of items because the adapted items may have weaker (or stronger) salience in the new setting.

In the unitary view of validity, evidence from content, criterion, and construct validity contribute to establishing confidence that the inferences and interpretations that one makes in a study are valid (Goodwin & Leach, 2003). The inferences that content-related evidence supports include the claim that a study's findings plausibly extend beyond the scale (the *implied* noumenon) to the theorized construct broadly (the *real* noumenon). As Hambleton (1984, p. 207) observe: "When the domain of items measuring an objective is unclear, only the weakest form of criterion-referenced test score interpretation is possible". Sireci (1998, p. 107) reaches a similar conclusion: "if the content of a test cannot be judged relevant to the construct measured, the validity of the empirical relationship between a test and its criterion is not defendable". Researchers assume a scale to be a viable sample of measures for a real noumenon. When they do not ratify that assumption, inferences generalized to a theoretical construct broadly may lack validity. Just as authors and reviewers limit claims to which findings generalize from a study sample to other populations, so too should authors use care when making generalizability claims from a scale to a theorized construct.

7 Conclusion

This paper reviews and synthesizes the content validity literature. We provide a comprehensive set of methods and quantitative metrics to support researchers who seek to evaluate and improve content validity of their survey scales. We demonstrate the proposed methods with a reflective survey scale example, then demonstrate efficacy of those methods in a field experiment. Researchers who develop new scales or adapt existing scales to a new research setting, can use the methods to establish confidence in content validity.

Acknowledgements

We thank the EIC, AE, and reviewers for their detailed and thoughtful feedback on prior versions of this paper. We also thank Richard Baskerville for kindly taking time for an extended discussion on generalizability and exposing us to various points of view on this ongoing and energetic debate in the IS community. These ideas were instrumental in guiding the recommendations that we make in this paper. We also thank Likoebe Maruping for facilitating a forum where we could present and test certain content

validity concepts, and Ed Rigdon for providing valuable feedback on our paper. Notwithstanding the assistance that these colleagues provided, we crafted the words and ideas in this paper and they remain our own. We take full responsibility for them. The J. Mack Robinson College of Business at Georgia State University supported this research.

References

Altman, N., & Krzywinski, M. (2015). Association, correlation and causation. *Nature Methods, 12*(10), 899-900.

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*(5), 732-740.

Avey, J. B., Avolio, B. J., Crossley, C. D., & Luthans, F. (2009). Psychological ownership: Theoretical extensions, measurement and relation to work outcomes. *Journal of Organizational Behavior, 30*(2), 173-191.

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly, 35*(2), 261-292.

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly, 27*(3), 459-489.

Baskerville, R. L., Kaul, M., & Storey, V.C. (2017). Establishing repeatability in design science research. In *Proceedings of the International Conference on Information Systems.*

Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly, 35*(2), 359-372.

Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly, 25*(1), 1-16.

Brown, R. D., & Hauenstein, N. M. (2005). Interrater agreement reconsidered: An alternative to the rwg indices. *Organizational Research Methods, 8*(2), 165-184.

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods, 15*(1), 17-32.

Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *MIS Quarterly, 33*(4), 689-707.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207-230.

Colquitt, J., Sabey, T., Rodell, J., & Hill, E. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*(10), 1243-1265.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194-197.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research, 3*(1), 60-95.

Delone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19*(4), 9-30.

DeLone, W. H., & McLean, E. R. (2016). Information systems success measurement. *Foundations and Trends in Information Systems, 2*(1), 1-116.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed., vol. 26). Los Angeles, CA: Sage.

Diamantopoulos, A. (2006). The error term in formative measurement models: Interpretation and modeling implications. *Journal of Modelling in Management, 1*(1), 7-17.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203-1218.

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management, 17*(4), 263-282.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105-124.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*(2), 370-388.

Ellwart, T., & Konradt, U. (2011). Formative versus reflective measurement: An illustration using work-family balance. *The Journal of Psychology, 145*(5), 391-417.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement*, *7*(1), 3-13.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50.

Franke, G. R., Preacher, K. J., & Rigdon, E. E. (2008). Proportional structural effects of formative indicators. *Journal of Business Research, 61*(12), 1229-1237.

Gable, G. G., Sedera, D., & Chan, T. (2008). Re-conceptualizing information system success: The IS-impact measurement model. *Journal of the Association for Information Systems, 9*(7), 377-408.

Gajewski, B. J., Coffland, V., Boyle, D. K., Bott, M., Price, L. R., Leopold, J., & Dunton, N. (2012). Assessing Content validity through correlation and relevance tools. *Methodology, 8*(3), 81-96.

Gefen, D. (2003). Assessing Unidimensionality through LISREL: An explanation and an example. *Communications of the Association for Information Systems, 12*, 23-47.

Gefen, D., & Straub, D. (2005). A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example. *Communications of the Association for Information Systems, 16*, 91-109.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology, 15*(4), 380-387.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*(2), 186-192.

Ghorpade, J., Lackritz, J., & Singh, G. (2006). Views of employee participation, higher order needs, altruism, pride in craftsmanship, and collectivism: Implications for organizational practice and public policy. *Journal of Applied Social Psychology, 36*(10), 2474-2491.

Goodwin, L. D., & Leech, N. L. (2003). Meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181-191.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Eductational Communication and Technology, 29*(2), 75-91

Guion, R. M. (1977). Content validity—the source of my discontent. *Applied Psychological Measurement, 1*(1), 1-10.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science, 40*(3), 414-433.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Beck (Ed.), *A guide to criterion referenced test construction.* Baltimore, MD: John Hopkins University Press.

Hardin, A. M., Chang, J. C.-J., & Fuller, M. A. (2008). Formative vs. reflective measurement: Comment on Marakas, Johnson, and Clay (2007). *Journal of the Association for Information Systems, 9*(9), 519-534.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology, 3*, 81-117.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Hemphill, J. K., & Westie, C. M. (1950). The measurement of group dimensions. *The Journal of Psychology, 29*(2), 325-342.

Hendrickson, A. R., Massey, P. D., & Cronan, T. P. (1993). On the test-retest reliability of perceived usefulness and perceived ease of use scales. *MIS Quarterly, 17*(2), 227-230.

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods, 2*(2), 175-186.

Hoehle, H., & Venkatesh, V. (2015). Mobile application usability: Conceptualization and instrument development. *MIS Quarterly, 39*(2), 435-472.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199-218.

Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., & Pollard, B. (2014). Discriminant content validity: A quantitative methodology for assessing content of theory-based measures, with illustrative applications. *British Journal of Health Psychology, 19*(2), 240-257.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276-2284.

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*(2), 161-167.

Krosnick, J. A., & Presser, S. (2018). Questionnaire design. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 439-455). Cham: Springer.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575.

Lebreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80-128.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information Systems Research*, *14*(3), 221-243.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, *16*(3), 294-304.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of CVI, T, r_{WG} (J), and $r^*_{WG(J)}$ indexes. *Journal of Applied Psychology, 84*(4), 640-647.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340-364.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly, 35*(2), 293-334.

Marakas, G. M., Johnson, R. D., & Clay, P. F. (2007). The evolving nature of the computer self-efficacy construct: An empirical investigation of measurement construction, validity, reliability and stability over time. *Journal of the Association for Information Systems, 8*(1), 16-46.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behavior, 23*(4), 311-318.

McMillan, S. S., King, M., & Tully, M. P. (2016). How to use the nominal group and Delphi techniques. *International Journal of Clinical Pharmacy, 38*, 655-662

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Mosser, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement,* 7, 191-205.

Nunnally, J. (1978). *Psychometric methods*. New York, NY: McGraw-Hill.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research, 1*(2), 66-82.

Parker, K. C., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103*(3), 367-373.

Petter, S., Straub, D. W., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623-656.

Pierce, J. L., Kostova, T., & Dirks, K. T. (2001). Toward a theory of psychological ownership in organizations. *Academy of Management Review, 26*(2), 298-310.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459-467.

Ringle, C. M., Sarstedt, M., & Straub, D. (2012). A critical look at the use of PLS-SEM in *MIS Quarterly*. *MIS Quarterly, 36*(1), iii-xii.

Robinson, M. A. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management, 57*(3), 739-750.

Rovinelli, R. J., & Hambleton, R. K. (1976). On the use of content specialists in the assessment of criterion-referenced test item validity. In *Proceedings of the 60th Annual Meeting of the American Educational Research Association*.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53-61.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management, 19*(2), 385-417.

Seddon, P. B., & Scheepers, R. (2015). Generalization in IS research: A critique of the conflicting positions of Lee & Baskerville and Tsang & Williams. *Journal of Information Technology*, 30(1), 30-43.

Segars, A. H. (1997). Assessing the unidimensionality of measurement: A paradigm and illustration within the context of information systems research. *Omega, 25*(1), 107-121.

Serrador, P., & Turner, R. (2015). The relationship between project success and project efficiency. *Project Management Journal, 46*(1), 30-39.

Shaft, T. M., Sharfman, M. P., & Wu, W. W. (2004). Reliability assessment of the attitude towards computers instrument. *Computers in Human Behavior, 20*(5), 661-689.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research,* 45, 83-117

Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1-19.

Straub, D., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13*, 380-427.

Tan, C.-W., Benbasat, I., & Cenfetelli, R. T. (2013). IT-mediated customer service content and delivery in electronic governments: An empirical investigation of the antecedents of service quality. *MIS Quarterly, 37*(1), 77-109.

Torkzadeh, G., & Doll, W. J. (1994). The test-retest reliability of user involvement instruments. *Information & Management, 26*(1), 21-31.

Tracy, J. L., & Robins, R. W. (2007). The psychological structure of pride: A tale of two facets. *Journal of Personality and Social Psychology, 92*(3), 506-525.

Trochim, W. M. K. (2006). Introduction to validity. *Web Center for Social Research Methods.* Retrieved from http://www.socialresearchmethods.net/kb/introval.php

Urbach, N., & Ahlemann, F. (2010). Structural equation modeling in information systems research using partial least squares. *Journal of Information Technology Theory and Application, 11*(2), 5-40.

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PloS One, 8*(9), 1-7.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327.

Wiliam, D. (1993). Validity, dependability and reliability in national curriculum assessment. *The Curriculum Journal,* 4(3), 335-350.

Williams, L. J., Hartman, N., & Cavazotte, F. (2010). Method variance and marker variables: A review and comprehensive CFA marker technique. *Organizational Research Methods, 13*(3), 477-514.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210.

Wolff, H.-G., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods, 37*(1), 48-58.

Wright, R. T., Campbell, D. E., Thatcher, J. B., & Roberts, N. H. (2012). Operationalizing multidimensional constructs in structural equation modeling: Recommendations for IS research. *Communications of the Association for Information Systems, 30*, 367-412.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442.

Appendix A: Current Validity Testing Practices in IS Literature

We identified papers in leading IS journals that used survey methods to collect data. Table A1 presents the full list of papers. Table A2 details content validity practices documented in each paper. We code papers that discuss content validity assessments without details as "face validity".

Table A1. IS Literature using Survey Methods

#	Year	Journal	Study using survey methods
1	2018	*MISQ*	Addas, S., & Pinsonneault, A. (2018). E-mail interruptions and individual performance: Is there a silver lining? *MIS Quarterly*, *42*(2), 381-406.
2	2018	*MISQ*	Moody, G. D., Siponen, M., & Pahnila, S. (2018). Toward a unified model of information security policy compliance. *MIS Quarterly*, *42*(1), 285-311.
3	2018	*MISQ*	Adjerid, I., Peer, E., & Acquisti, A. (2018). Beyond the privacy paradox: Objective versus relative risk in privacy decision making. *MIS Quarterly*, *42*(2), 465-488.
4	2018	*MISQ*	Ye, H., & Kankanhalli, A. (2018). User service innovation on mobile phone platforms: Investigating impacts of lead userness, Toolkit support, and design autonomy. *MIS quarterly*, *42*(1), 165-187.
5	2018	*MISQ*	Benitez, J., Ray, G., & Henseler, J. (2018). Impact of information technology infrastructure flexibility on mergers and acquisitions. *MIS Quarterly*, *42*(1), pp.25-43.
6	2018	*MISQ*	Retana, G. F., Forman, C., Narasimhan, S., Niculescu, M. F., & Wu, D. J. (2018). Technology support and post-adoption IT service use: Evidence from the cloud. *MIS Quarterly*, *42*(3), 961-978.
7	2018	*MISQ*	Ho, S. Y., & Lim, K. H. (2018). Nudging moods to induce unplanned purchases in imperfect mobile personalization contexts. *MIS Quarterly*, *42*(3), 757-778.
8	2018	*MISQ*	Avgar, A., Tambe, P., & Hitt, L. M. (2018). Built to learn: How work practices affect employee learning during healthcare information technology implementation. *MIS Quarterly*, *42*(2), 645-660.
9	2018	*MISQ*	Romanow, D., Rai, A., & Keil, M. (2018). CPOE-enabled coordination: Appropriation for deep structure use and impacts on patient outcomes. *MIS Quarterly*, *42*(1), 189-212.
10	2018	*MISQ*	Li, X., & Wu, L. (2018). Herding and social media word-of-mouth: Evidence from Groupon. *Management Information Systems Quarterly*, *42*(4), 1331-1351.
11	2018	*MISQ*	Daniel, S. L., Maruping, L. M., Cataldo, M., & Herbsleb, J. (2018). The impact of ideology misfit on open source software communities and companies. *MIS Quarterly*, *42*(4), 1069-1096.
12	2018	*MISQ*	Chen, A., & Karahanna, E. (2018). Life interrupted: The effects of technology-mediated work interruptions on work and nonwork outcomes. *MIS Quarterly*, *42*(4), 1023-1042.
13	2018	*MISQ*	Thatcher, J. B., Wright, R. T., Sun, H., Zagenczyk, T. J., & Klein, R. (2018). Mindfulness in information technology use: Definitions, distinctions, and a new measure. *MIS Quarterly*, *42*(3), 831-848.
14	2018	*MISQ*	Srivastava, S. C., & Chandra, S. (2018). Social presence in virtual world collaboration: An uncertainty reduction perspective using a mixed methods approach. *MIS Quarterly*, *42*(3), 779-804.
15	2018	*MISQ*	Karahanna, E., Xu, S. X., Xu, Y., & Zhang, N. A. (2018). The needs-affordances-features perspective for the use of social media. *MIS Quarterly*, *42*(3), 737-756.
16	2018	*MISQ*	Valacich, J. S., Wang, X., & Jessup, L. M. (2018). Did I buy the wrong gadget? How the evaluability of technology features influences technology feature preferences and subsequent product choice. *MIS Quarterly*, *42*(2), 633-644.
17	2018	*MISQ*	Ye, S., Viswanathan, S., & Hann, I. H. (2018). The value of reciprocity in online barter markets: An empirical investigation. *MIS Quarterly*, *42*(2), 521-549.
18	2018	*MISQ*	Gunarathne, P., Rui, H., & Seidmann, A. (2018). When social media delivers customer service: Differential customer treatment in the airline industry. *MIS Quarterly*, *42*(2), 489-520.
19	2018	*MISQ*	Vance, A., Jenkins, J. L., Anderson, B. B., Bjornn, D. K., & Kirwan, C. B. (2018). Tuning out security warnings: A longitudinal examination of habituation through fMRI, eye tracking, and field experiments. *MIS Quarterly*, *42*(2), 355-380.

Table A1. IS Literature using Survey Methods

20	2018	*ISR*	Anderson, E. G., Jr., Chandrasekaran, A., Davis-Blake, A., & Parker, G. G. (2018). Managing distributed product development projects: Integration strategies for time-zone and language barriers. *Information Systems Research*, *29*(1), 42-69.
21	2018	*ISR*	Kim, H. W., Kankanhalli, A., & Lee, S. H. (2018). Examining gifting through social network services: A social exchange theory perspective. *Information Systems Research*, *29*(4), 805-828.
22	2018	*ISR*	Tiwana, A. (2018). Platform synergy: Architectural origins and competitive consequences. *Information Systems Research*, *29*(4), 829-848.
23	2018	*ISR*	Robert, L. P., Jr., Dennis, A. R., & Ahuja, M. K. (2018). Differences are different: Examining the effects of communication media on the impacts of racial and gender diversity in decision-making teams. *Information Systems Research*, *29*(3), 525-545.
24	2018	*ISR*	Venkatesh, V., Rai, A., & Maruping, L. M. (2018). Information systems projects and individual developer outcomes: Role of project managers and process control. *Information systems research*, *29*(1), 127-148.
25	2018	*JMIS*	Ma, X., Khansa, L., & Kim, S. S. (2018). Active community participation and crowdworking turnover: A longitudinal model and empirical test of three mechanisms. *Journal of Management Information Systems*, *35*(4), 1154-1187.
26	2018	*JMIS*	Magni, M., Ahuja, M. K., & Maruping, L. M. (2018). Distant but fair: Intra-team justice climate and performance in dispersed teams. *Journal of Management Information Systems*, *35*(4), 1031-1059.
27	2018	*JMIS*	Hashim, M. J., Kannan, K. N., & Wegener, D. T. (2018). Central role of moral obligations in determining intentions to engage in digital piracy. *Journal of Management Information Systems*, *35*(3), 934-963.
28	2018	*JMIS*	Kathuria, A., Mann, A., Khuntia, J., Saldanha, T. J., & Kauffman, R. J. (2018). A strategic value appropriation path for cloud computing. *Journal of Management Information Systems*, *35*(3), 740-775.
29	2018	*JMIS*	Havakhor, T., & Sabherwal, R. (2018). Team processes in virtual knowledge teams: The effects of reputation signals and network density. *Journal of Management Information Systems*, *35*(1), 266-318.
30	2018	*JAIS*	Hsu, J., Chiu, C. M., Lowry, P. B., & Liang, T. P. (2017). Solving the interpretational-confounding and interpretational-ambiguity problems of formative construct modeling in behavioral research: proposing a two-stage fixed-weight redundancy approach. *Journal of the Association for Information Systems, 19*(7), 618-671.
31	2018	*JAIS*	Harrison, A. (2018). The effects of media capabilities on the rationalization of online consumer fraud. *Journal of the Association for Information Systems*, *19*(5), 408-440.
32	2018	*JAIS*	Tams, S., Thatcher, J. B., & Grover, V. (2018). Concentration, competence, confidence, and capture: An experimental study of age, interruption-based technostress, and task performance. *Journal of the Association for Information Systems*, *19*(9), 857-908.
33	2018	*JAIS*	Tarafdar, M., & Tanriverdi, H. (2018). Impact of the information technology unit on information technology-embedded product innovation. *Journal of the Association for Information Systems*, *19*(8), 716-751.
34	2018	*JAIS*	Choi, B., Wu, Y., Yu, J., & Land, L. (2018). Love at first sight: The interplay between privacy dispositions and privacy calculus in online social connectivity management. *Journal of the Association for Information Systems*, *19*(3), 124-151.
35	2019	*MISQ*	Wunderlich, P., Veit, D. J., & Sarker, S. (2019). Adoption of sustainable technologies: A mixed-methods study of German households. *MIS Quarterly*, *43*(2), 673-691.
36	2019	*MISQ*	Moeini, M., & Rivard, S. (2019). Responding—or not—to information technology project risks: an integrative model. *MIS Quarterly*, *43*(2), 475-500.
37	2019	*MISQ*	Liang, H., Xue, Y., Pinsonneault, A., & Wu, Y. (2019). What users do besides problem-focused coping when facing it security threats: An emotion-focused coping perspective. *MIS Quarterly*, *43*(2), 373-394.
38	2019	*MISQ*	Maruping, L. M., Daniel, S. L., & Cataldo, M. (2019). Developer centrality and the impact of value congruence and incongruence on commitment and code contribution activity in open source software communities. *MIS Quarterly*, *43*(3), 951-976.

Table A1. IS Literature using Survey Methods

39	2019	*MISQ*	Venkatesh, V., Sykes, T., Chan, F. K., Thong, J. Y., & Hu, P. J. (2019). Children's internet addiction, family-to-work conflict, and job outcomes: A study of parent-child dyads. *MIS Quarterly*, *43*(3), 903-927.
40	2019	*MISQ*	Wu, J., Huang, L., & Zhao, J. L. (2019). Operationalizing regulatory focus in the digital age: Evidence from an e-commerce context. *MIS Quarterly*, *43*(3), 745-764.
41	2019	*MISQ*	Kim, A., & Dennis, A. R. (2019). Says Who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, *43*(3), 1025-1039.
42	2019	*MISQ*	Nishant, R., Srivastava, S. C., & Teo, T. S. (2019). Using polynomial modeling to understand service quality in e-government websites. *MIS Quarterly*, *43*(3), 807-826.
43	2019	*MISQ*	Geva, H., Oestreicher-Singe, G., & Saar-Tsechansky, M. (2019). Using retweets when shaping our online persona: Topic modeling approach. *MIS Quarterly*, *43*(2), 501-524.
44	2019	*MISQ*	Bapna, S., Benner, M. J., & Qiu, L. (2019). Nurturing online communities: An empirical investigation. *MIS Quarterly*, *43*(2), 425-452.
45	2019	*MISQ*	James, T. L., Wallace, L., & Deane, J. K. (2019). Using organismic integration theory to explore the associations between users' exercise motivations and fitness technology feature set use. *MIS Quarterly*, *43*(1), 287-312.
46	2019	*MISQ*	Burtch, G., & Chan, J. (2019). Investigating the relationship between medical crowdfunding and personal bankruptcy in the United States: Evidence of a digital divide. *MIS Quarterly*, *43*(1), 237-262.
47	2019	*ISR*	Bouayad, L., Padmanabhan, B., & Chari, K. (2019). Audit policies under the sentinel effect: Deterrence-driven algorithms. *Information Systems Research*, *30*(2), 466-485.
48	2019	*ISR*	Yang, M., Ren, Y., & Adomavicius, G. (2019). Understanding user-generated content and customer engagement on Facebook business pages. *Information Systems Research*, *30*(3), 839-855.
49	2019	*ISR*	Buckman, J. R., Bockstedt, J. C., & Hashim, M. J. (2019). Relative privacy valuations under varying disclosure characteristics. *Information Systems Research*, *30*(2), 375-388.
50	2019	*ISR*	Koh, T. K. (2019). Adopting seekers' solution exemplars in crowdsourcing ideation contests: Antecedents and consequences. *Information Systems Research*, *30*(2), 486-506.
51	2019	*ISR*	Crossler, R. E., & Bélanger, F. (2019). Why would I use location-protective settings on my smartphone? Motivating protective behaviors and the existence of the privacy knowledge-belief gap. *Information Systems Research*, *30*(3), 995-1006.
52	2019	*ISR*	Lee, J. S., Keil, M., & Shalev, E. (2019). Seeing the trees or the forest? The effect of IT project managers' mental construal on IT project risk management activities. *Information Systems Research*, *30*(3), 1051-1072.
53	2019	*ISR*	Wang, W., & Wang, M. (2019). Effects of sponsorship disclosure on perceived integrity of biased recommendation agents: psychological contract violation and knowledge-based trust perspectives. *Information Systems Research*, *30*(2), 507-522.
54	2019	*JMIS*	Chan, T. K., Cheung, C. M., & Wong, R. Y. (2019). Cyberbullying on social networking sites: The crime opportunity and affordance perspectives. *Journal of Management Information Systems*, *36*(2), 574-609.
55	2019	*JMIS*	Khuntia, J., Kathuria, A., Saldanha, T. J., & Konsynski, B. R. (2019). Benefits of IT-enabled flexibilities for foreign versus local firms in emerging economies. *Journal of Management Information Systems*, *36*(3), 855-892.
56	2019	*JMIS*	Pirkkalainen, H., Salo, M., Tarafdar, M., & Makkonen, M. (2019). Deliberate or instinctive? Proactive and reactive coping for technostress. *Journal of Management Information Systems*, *36*(4), 1179-1212.
57	2019	*JMIS*	Phang, C. W., Luo, X., & Fang, Z. (2019). Mobile time-based targeting: Matching product-value appeal to time of day. *Journal of Management Information Systems*, *36*(2), 513-545.
58	2019	*JMIS*	Hu, Y., Xu, A., Hong, Y., Gal, D., Sinha, V., & Akkiraju, R. (2019). Generating business intelligence through social media analytics: Measuring brand personality with consumer-, employee-, and firm-generated content. *Journal of Management Information Systems*, *36*(3), 893-930.

Table A1. IS Literature using Survey Methods

59	2019	*JMIS*	Klier, J., Klier, M., Thiel, L., & Agarwal, R. (2019). Power of mobile peer groups: A design-oriented approach to address youth unemployment. *Journal of Management Information Systems*, *36*(1), 158-193.
60	2019	*JMIS*	Wang, K., & Nickerson, J. V. (2019). A Wikipedia-based method to support creative idea generation: The role of stimulus relatedness. *Journal of Management Information Systems*, *36*(4), 1284-1312.
61	2019	*JMIS*	Oshri, I., Dibbern, J., Kotlarsky, J., & Krancher, O. (2019). An information processing view on joint vendor performance in multi-sourcing: The role of the guardian. *Journal of Management Information Systems*, *36*(4), 1248-1283.
62	2019	*JMIS*	Lowry, P. B., Zhang, J., Moody, G. D., Chatterjee, S., Wang, C., & Wu, T. (2019). An integrative theory addressing cyberharassment in the light of technology-based opportunism. *Journal of Management Information Systems*, *36*(4), 1142-1178.
63	2019	*JMIS*	Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, *36*(3), 931-968.
64	2019	*JMIS*	Steffen, J. H., Gaskin, J. E., Meservy, T. O., Jenkins, J. L., & Wolman, I. (2019). Framework of affordances for virtual reality and augmented reality. *Journal of Management Information Systems*, *36*(3), 683-729.
65	2019	*JMIS*	Winkler, T. J., & Wulf, J. (2019). Effectiveness of it service management capability: Value co-creation and value facilitation mechanisms. *Journal of Management Information Systems*, *36*(2), 639-675.
66	2019	*JMIS*	Yuan, L., & Dennis, A. R. (2019). Acting like humans? Anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems*, *36*(2), 450-477.
67	2019	*JMIS*	Dunn, B. K., Ramasubbu, N., Galletta, D. F., & Lowry, P. B. (2019). Digital borders, location recognition, and experience attribution within a digital geography. *Journal of Management Information Systems*, *36*(2), 418-449.
68	2019	*JMIS*	Craig, K., Thatcher, J. B., & Grover, V. (2019). The IT identity threat: A conceptual definition and operational measure. *Journal of Management Information Systems*, *36*(1), 259-288.
69	2019	*JMIS*	Maruping, L. M., Venkatesh, V., Thong, J. Y., & Zhang, X. (2019). A risk mitigation framework for information technology projects: A cultural contingency perspective. *Journal of Management Information Systems*, *36*(1), 120-157.
70	2019	*JAIS*	Teubner, T., & Flath, C. M. (2019). Privacy in the sharing economy. *Journal of the Association for Information Systems*, *20*(3), 213-242.
71	2019	*JAIS*	Zhang, Y., Lu, T., Phang, C. W., & Zhang, C. (2019). Scientific knowledge communication in online Q&A communities: Linguistic devices as a tool to increase the popularity and perceived professionalism of knowledge contributions. *Journal of the Association for Information Systems*, *20*(8), 1129-1173.
72	2019	*JAIS*	Morana, S., Kroenung, J., Maedche, A., & Schacht, S. (2019). Designing process guidance systems. *Journal of the Association for Information Systems*, *20*(5), 499-535
73	2019	*JAIS*	Stich, J. F., Tarafdar, M., Stacey, P., & Cooper, C. (2019). Appraisal of email use as a source of workplace stress: A person-environment fit approach. *Journal of the Association for Information Systems*, *20*(2), 132-160.
74	2019	*JAIS*	Tam, K. Y., Feng, K. Y., & Kwan, S. (2019). The role of morality in digital piracy: Understanding the deterrent and motivational effects of moral reasoning in different piracy contexts. *Journal of the Association for Information Systems*, *20*(5), 604-628.
75	2019	*JAIS*	Wortmann, F., Thiesse, F., & Fleisch, E. (2019). The impact of goal-congruent feature additions on core IS feature use: When more is less and less is more. *Journal of the Association for Information Systems*, *20*(7), 953-985.
76	2019	*JAIS*	Gerlach, J. P., Buxmann, P., & Dinev, T. (2019). "They're all the same!" Stereotypical thinking and systematic errors in users' privacy-related judgments about online services. *Journal of the Association for Information Systems*, *20*(6), 787-823.
77	2019	*JAIS*	Lin, S., & Armstrong, D. J. (2019). Beyond information: The role of territory in privacy management behavior on social networking sites. *Journal of the Association for Information Systems*, *20*(4), 434-475.

Table A1. IS Literature using Survey Methods

78	2019	*JAIS*	Kwak, D. H., Holtkamp, P., & Kim, S. S. (2019). Measuring and controlling social desirability bias: Applications in information systems research. *Journal of the Association for Information Systems*, *20*(4), 317-345.

Table A2. Validity Tests

	Construct validity						Content validity								
	Indicator reliability	Internal consistency	Convergent validity	Discriminant validity	Indicator weight	Collinearity	Construct clarity	Item clarity	Relevance (internal congruence)	Corruption (external congruence)	Indicator sufficiency	Indicator parsimony	Content consistency	Content stability	Face validity
1	L	α	AVE	XL, √AVE		VIF			Sort, p_{sa}					TrT	
2	L	α	AVE	XL, $\chi^2\Delta$											Pn, Pi
3	L	α, ρc	AVE	XL, √AVE											
4															
5	L	α, ρc	AVE	XL, √AVE	W	VIF									Pn
6	L			XL	W	VIF									Pn
7															
8	L		AVE	XL, √AVE											
9		α													
10	L	α, ρc	AVE	XL, √AVE		VIF									Pn
11		α													
12	L	α	AVE	XL, √AVE	W	VIF									
13	L	ρc	AVE	√AVE					Sort	Sort					Pn, Pi
14	L	α, ρc	AVE	XL, $\chi^2\Delta$, √AVE		VIF			Sort, p_{sa}	Sort			α, ρc	TrT	
15	L	α	AVE	XL, √AVE											Pn, Pi
16	L	α, ρc	AVE	XL, √AVE											
17		α													
18	L	α		XL											
19		α				VIF									
20															
21		α				VIF									
22	L	α, ρc	AVE	XL, √AVE					Sort, $p_{sa,}$ kappa						Pn
23	L	α	AVE	XL											
24	L	α				VIF									
25	L	α, ρc	AVE	$\chi^2\Delta$, √AVE											

Table A2. Validity Tests

26	L	ρc	AVE	$\chi^2\Delta$											
27		ρc		$\chi^2\Delta$										TrT	
28	L	α, ρc	AVE	XL, √AVE											
29	L	α, ρc	AVE	XL, HMr	W	VIF							α, ρc	TrT	Pn
30															
31	L	α, ρc	AVE	XL, √AVE											
32	L	α	AVE	XL, √AVE					Sort						Pn, Pi
33	L	α, ρc	AVE	XL, √AVE									α		Pi
34	L	α, ρc	AVE	XL, $\chi^2\Delta$, √AVE											Pn
35	L	α, ρc	AVE	XL, √AVE											
36	L	α, ρc	AVE	√AVE		VIF		Pn	Sort, kappa				α, ρc		Pn, Pi
37	L	α, ρc	AVE	XL, √AVE				Pn	Sort, kappa, IR				α, ρc		Pn, Pi
38	L	α	AVE	XL, √AVE		VIF			Sort, p_{sa}	Sort XL			α	TrT	Pn, Pi
39	L	α		XL, √AVE											
40	L	α, ρc	AVE	XL, √AVE	W	VIF		LX	Sort						Pi
41								LX							Pn
42		α													
43	L	α, ρc	AVE	XL, √AVE		VIF									
44															
45									Sort, p_{sa}						Pn
46	L	α, ρc	AVE	XL, √AVE		VIF									Pn
47															
48															Pn
49															
50															
51															
52	L	α, ρc	AVE	XL, √AVE											Pn, Pi
53	L	α	AVE												
54	L	α		XL, √AVE											

Table A2. Validity Tests

55	L	α	AVE	XL, $\chi^2\Delta$				Rating	Sort & rating kappa				α		Pn, Pi
56	Pi	α	Pi	Pi		VIF		Pn					α		Pn, Pi
57	L	ρc	AVE	√AVE											
58		α, ρc	AVE	√AVE					Sort						Pn
59	L														
60	L	α, ρc	AVE	XL, √AVE											
61		α													
62	L	ρc	AVE	XL, √AVE				Pn							Pn
63	L	α	AVE	√AVE, $\chi^2\Delta$		VIF									
64		α													
65															
66	L	α		XL		VIF							α		Pn, Pi
67	L	α													
68	L	α	AVE	XL											
69	L	α, ρc	AVE	XL, √AVE		VIF			Sort				α, ρc		Pn, Pi
70	L	ρc	AVE	√AVE	W	VIF									
71	L	α, ρc	AVE	HTMT											Pn
72															
73	L	α, ρc	AVE	XL, √AVE				LX							Pn
74	L	α, ρc	AVE	√AVE											
75	L	α, ρc	AVE	XL, √AVE				Pi							Pi
76	L	α, ρc	AVE	XL, √AVE									α		Pi
77	L	α, ρc	AVE	XL, √AVE	W	VIF			Sort				α, ρc		Pn, Pi
78	L	α, ρc	AVE	XL, √AVE, $\chi^2\Delta$	W	VIF									Pn, Pi
79	L	α, ρc	AVE	√AVE				Pn							Pn

Note: L: loading, α: Cronbach's coefficient alpha tau-equivalent reliability, ρc: rho C composite reliability, AVE: average variance extracted > 0.5, √AVE: Fornell-Larcker criterion, XL: cross-loading, $\chi^2\Delta$: Chi-squared difference test, W : formative indicator weights, VIF: variance inflation factor < 10, p_{sa}: proportion of substantive agreement, TrT : test-retest reliability, Pn: pre-test panel with some content validity consideration, Pi: pilot study with content validity consideration, LX: double language translation.

Appendix B: External Congruence and Unidimensionality

External congruence shares some properties with the more familiar unidimensionality concept often associated with construct validity. Unidimensionality requires "the existence of a single trait or construct underlying a set of measures" (Gerbing & Anderson, 1988, p. 186). Unidimensional scales measure a single trait (Segars, 1997). Researchers commonly employ several statistical approaches for such validity considerations. With structural equation modeling, the desire to assess unidimensionality for study data has led researchers to redefine unidimensionality with a mathematical definition (Gerbing & Anderson, 1988). Two criteria emerging from this definition include internal consistency and external consistency.

Internal consistency exists when a common factor's indicators highly correlate with one another, whereas external consistency exists when indicators have a low correlation to indicators of other constructs in a model. One calculates correlations for all item pairs and between an item and each latent construct. Internal consistency exists when a pair of items (ρ_{ij}) that should inform a single target construct (ξ) have the same correlation with each other (indicators i and j) as the product of their individual correlations to the latent construct($\rho_{i\xi}$) (i.e., $\rho_{ij} = \rho_{i\xi} \ \rho_{j\xi}$) (Gerbing & Anderson, 1988). Internal consistency exists when the only correlation between two measurement items is a function of the correlation of their common latent variable. Internal consistency is established by demonstrating that non-common variance (potential multi-dimensionality or relationship through a second latent factor) does not contribute significantly to the correlation between the two items (Gefen 2003).

External consistency exists when any correlation between two items on different scales result from the relationship that those items have with their intended latent constructs and the subsequent correlation between those two latent variables. Here, the common variance is a function of the correlation between two factors (indicators i and p), the correlation of an item and its intended factor ($\rho_{i\xi}$), and a second item ($\rho_{p\xi^*}$) and that second item's intended factor ($\rho_{\xi\xi^*}$) (i.e., $\rho_{ip} = \rho_{i\xi} \ \rho_{\xi\xi^*} \ \rho_{p\xi^*}$) (Gerbing & Anderson, 1988). External consistency exists when the only correlation between two measurement items is a function of the correlation of their respective latent variables. External consistency is established by demonstrating that non-common variance (potential multi-dimensionality or relationship through additional latent factors) does not contribute significantly to the correlation between those two items (Gefen, 2003).

Internal and external consistency relate to convergent validity and discriminant validity, respectively. Convergent validity reflects the extent to which two items capture a common construct (Carlson & Herdman, 2012). Researchers commonly calculate the Pearson's correlation coefficient (r) with support for convergent validity appearing as low as $r = 0.40$ and strong support appearing at $r \geq 0.70$. A somewhat more rigorous hypothesis test involves examining the ratio of factor loadings to their respective standard errors (Segars, 1997). One demonstrates convergent validity when all measurement items load on their intended latent construct with a significant t-value (Gefen & Straub, 2005). A popular alternate metric for convergent validity involves calculating the average variance in the indicators that the intended construct accounts for. One calculates an average variance extracted (AVE) metric by averaging the squared standardized factor loading for the indicators of a target latent construct. An AVE value greater than 0.50 represents a good indicator that the latent construct accounts for the majority of variance in the indicators and, thus, supports convergent validity (MacKenzie et al., 2011).

Discriminant validity reflects the extent to which measures of distinct concepts differ (Bagozzi & Phillips, 1982). Researchers need to show that one construct's indicators differ from indicators of other constructs (MacKenzie et al., 2011). Researchers often use cross loadings to identify poor discriminant validity. One calculates pearson's correlation coefficient (r) for each pair of items. Items should load highest on their intended construct (Hair et al., 2012). In addition, any correlations above $r = 0.2$ among items that should belong to different latent constructs may suggest poor discriminant validity (Robinson, 2018). The Fornell and Larcker (1981) criterion represents a common approach to determine if a latent variable shares more variance with its associated indicators than it shares with other constructs in the same model. One calculates the criterion by calculating the AVE measures for all constructs and the squared correlation (r) between constructs. Constructs in a pair with AVE measures that exceed the squared correlation between them evidence discriminant validity (MacKenzie et al., 2011; Segars, 1997).

Convergent and discriminant validity expose common variance, but do not encompass non-common variance that is central to the definition of unidimensionality. As a result, these measures cannot directly establish unidimensionality (Geffen, 2003; Segars, 1997). Other factor analytic techniques, such as principle component analysis (PCA), have similar usefulness as scale-development tools that can detect severe unidimensionality problems, but are not a robust test for unidimensionality (Gerbing & Anderson,

1988). Figure B1 depicts the nesting of validity measures. The innermost metrics address only a subset of considerations for each form of validity. For example, convergent and discriminant validity metrics inform but do not establish unidimensionality. As a separate consideration, one cannot assess convergent and discriminant validity of formative constructs using the mathematical method discussed above because one does not expect correlation among formative items (Hair et al., 2012).

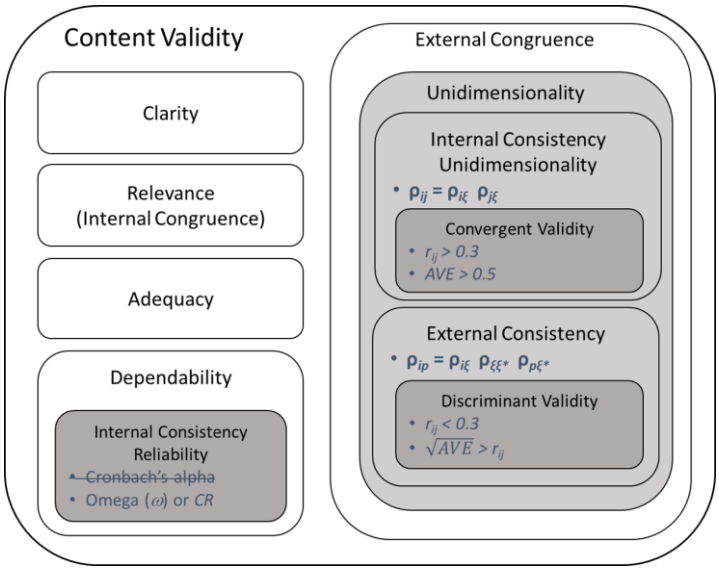

Figure B1. External Congruence and Unidimensionality

Just as convergent and discriminant validity inform, but cannot establish, unidimensionality, unidimensionality tests can inform, but cannot establish, external congruence. Although external congruence and unidimensionality appear to have similar definitions, operationalized unidimensionality falls short as a suitable method to establish external congruence. The universe of phenomena that describe the human condition is quite large. When researchers include an indicator for an orbiting construct in a study, their data will include measurement of phenomenon they did not intent. All metrics that use study data examine only the constellation of constructs in the operationalized model. They cannot consider the possibility of other unspecified orbiting constructs. As a result, unidimensionality metrics remain blind to misspecification. Multidimensional items that inform three or more orbiting constructs can have good internal and external consistency when researchers limit their validity assessment to theorized constructs. All metrics that researchers calculate on study data remain blind to the possibility of orbiting constructs that the study's model does not operationalize.

Traditional unidimensionality tests incorrectly assume that all plausibly relevant latent factors are included in external consistency tests during a study's main assessment. IS scholars who access discriminant validity should include "measures of similar constructs that are potential confounded with the focal construct" (MacKenzie et al., 2011, p. 311). However, in practice, one cannot easily do so given the infinite number of orbiting constructs that exist in the universe of the human condition. The scope of both discriminant validity and unidimensionality is limited to only theorized latent constructs. This blinds researchers from other orbiting constructs that may contaminate the latent construct's inferred properties. Unidimensionality indicates that measures are relatively clean with respect to the other measures in the main study, but it remains a poor indicator at best.

More importantly, a misspecification error does not simply constitute a random measurement error—it is systemic error due to a hidden covariate consistently appearing in the data. Such errors result in unreliable or erroneous conclusions (Segars, 1997). As Bagozzi & Phillips, 1982 (p. 460) note, "Operationally, this means that the estimates may confirm a relationship where in fact none exists, or they may mask an actual underlying relationship".

Appendix C: Survey Instruments Used in Field Experiment

We used six constructs and associated items in the experiment. We divided the constructs into two sets of constructs, which we called cluster A and cluster B.

Cluster A involved a set of orbiting constructs with scales appearing in published studies. We adapted self-efficacy (SE) from Avey, Avolio, Crossley, and Luthans (2009), pride in craftsmanship (PC) from Ghorpade, Lackritz, and Singh (2006), and authentic price (AP) from Tracy and Robins (2007). While these researchers previously validated these scales for a specific setting, their adaptation and application to a new setting undermines confidence in content validity (Haynes et al. 1995). Whereas SE and PiC represent stable traits, AP is a cognitive state that depends on the situation. As a result, study data does not represent a suitable option for evaluating dependability.

In the experiment, we prepared survey questions to examine behaviors and attitudes of college students toward work-products generated during assignments. As a result, we adapted all questions to the context of student subjects and classroom assignments.

In Tables C1 to C6, dirty constructs refer to constructs that we adapted from prior publication. We subjected the scales to several rounds of panel review using the item-rating methods that we describe in this paper for clarity, congruence, and dependability. This experiment did not involve any formative scales, so we did not employ the construct validity processes for adequacy. After collecting and analyzing final-study data we removed items to improve construct validity of the dirty item scale. The items we removed are noted with strikethrough in the tables below. We removed clean items (also marked with a strikethrough) following the final panel review and before we collected the primary study data.

Cluster A

Self-efficacy: a belief that I can successfully do specific tasks. For the purposes of this study, class assignments (including the work, the file & quiz answers created as part of that assignment) are the asks of interest.

This study is asking a series of questions to determine if the student (the person answering the survey) has confidence they can successfully do class assignments.

Table C1. Self-efficacy

Construct	Item
Dirty	DSE1: I am confident in my ability to contribute to my team's success on assignment work in this class
Dirty	DSE2: I can make a positive difference on assignment work for my collaboration team in this class
Dirty	DSE3: I am confident pushing high performance goals on assignment work for my collaboration team in this class
Clean	CSE1: I am confident in my ability to complete assignments in this class
Clean	CSE2: I am confident setting high goals in this class
Clean	CSE3: In this class I have the feeling I can handle the difficult situations
Note: questions answered using six-point scale from strongly disagree to strongly agree.	

Pride in craftsmanship: a general work ethic that exists to a greater (or lesser) extent in everyone. Individuals with a strong pride in craftsmanship believe that everyone should produce quality product, regardless of their likes or dislikes, presence (or absence) of instructors and aides, and should take pride in the results of their efforts.

This study is asking a series of questions to determine if the student (the person answering the survey) shares the belief that all students (and all people) should do quality work regardless of circumstances.

Table C2. Pride in Craftsmanship

Construct	Item
Dirty	DPC1: A student should do a decent job whether or not his/her teacher is around
Dirty	DPC2: A student should feel a sense of pride in his/her work
~~Dirty~~	~~DPC3: There is nothing wrong with doing a poor job on assignments if a person can get away with it~~
Dirty	DPC4: Regardless of whether a task is mental or manual, pleasant or unpleasant, it should be performed with the best of one's effort
Dirty	DPC5: Even if you dislike your assignment, you should do your best
Dirty	DPC6: Both in class and outside of class, I take pride in the quality of my work
Clean	CPC1: All students should do a decent job even when the teacher is not present
~~Clean~~	~~CPC2: I take pride in the work I do outside of school~~
~~Clean~~	~~CPC3: There is nothing wrong with doing a poor job if a person can get away with it~~
Clean	CPC4: Regardless of whether a task is mental or manual, pleasant or unpleasant, it should be performed with the best of one's effort
Clean	CPC5: Even if you dislike your assignment, you should do your best
Clean	CPC6: Everyone should have pride in the quality of their work in all situations
Note: questions answered using six-point scale from strongly disagree to strongly agree.	

Authentic pride: pride associated with specific accomplishments bringing genuine feelings of self-worth. For the purposes of this study each assignment (including the work, the file & quiz answers created as part of that assignment) is an accomplishment.

This study is asking a series of questions to determine if the student (the person answering the survey) has pride in the assignments (the work, the file, and answers) they submit in class.

Table C3. Authentic Pride

Construct	Item	
Dirty	DAP1	My work on assignments in this class makes me feel accomplished
Dirty	DAP2	My work on assignments in this class makes me feel successful
Dirty	DAP3	My work on assignments in this class makes me feel like I am achieving
Dirty	DAP4	My work on assignments in this class makes me feel productive
Dirty	DAP5	I feel proud when I complete assignments in this class
Dirty	DAP6	I feel proud of my assignment scores in this class
Dirty	DAP7	My work on assignments in this class makes me feel like I have self-worth
Dirty	DAP8	My work on assignments in this class makes me feel confident
Dirty	DAP9	When other students view my assignment files, I feel proud of my work
~~Dirty~~	~~DAP10~~	~~When other students use my ideas, I feel good about my role in the team~~
~~Dirty~~	~~DAP11~~	~~When other students benefit from the work I do on assignments, I am proud of my role in the class.~~
Clean	CAP1	Completing assignments in this class is a real accomplishment
Clean	CAP2	Completing assignments in this class makes me feel successful
Clean	CAP3	Completing assignments in this class makes me feel like I am achieving
Clean	CAP4	Completing assignments in this class is fulfilling
Clean	CAP5	Completing assignments in this class makes me feel productive
Clean	CAP6	I feel proud when I complete assignments in this class
~~Clean~~	~~CAP7~~	~~I feel proud of my assignment scores in this class~~
Note: questions answered using six-point scale from strongly disagree to strongly agree.		

Cluster B

Cluster B involves newly developed scales related to the concept of psychological ownership. psychological ownership (Pierce, Kostova, & Dirks, 2001) has been widely used to capture pride relative to a place (such as the workplace or residence). Adapting existing scales for psychological ownership of place to psychological ownership of an object involved a very significant contextual change. Existing place-oriented scales draw strongly on the idea of "having a place" and "possessing territory and space" (Pierce et al., 2001, p. 300). To better align with objects, we chose to devise entirely new scales for each component motivated by the theorized "routes" to psychological ownership: controlling the target, coming to intimately know the target, and investing the self into the target.

Instead of working with previously existing scales, constellation B involves new reflective scales that constitute a second order formative construct. We theorized three dimensions (control, knowledge, and self-identity) and developed new reflective scales to measure each. As a result, the construct validity challenges with these orbiting constructs represent a different challenge than the constructs in cluster A.

Control: feelings of ownership arise as the individual is successful in experiencing control over an object. The object of interest in this study is the assignment (the ideas, the work that demonstrates those ideas and the file itself). Ownership means *the ability to use and to control the use* of objects. This includes not only how the student uses their own work (for studying or guiding their work on later tasks), but also knowingly allowing the ideas/work/file to come into contact with other people (students, aids, instructor, LMS) under circumstances the owner controls.

This study is asking a series of questions to determine if the student (the person answering the survey) has a feeling of control over their assignment work (the ideas and the assignment file).

Table C4. Psychological Ownership: Control

Construct	Item
~~Dirty~~	~~DOC1: I control the work I submit for assignments in this class~~
Dirty	DOC2: I determine who can use the work I do for this class
Dirty	DOC3: I regulate who can access the work I do for this class
Dirty	DOC4: I determine when others can access the work I do for this class
Dirty	DOC5: I control where others can use the work I do for this class
Dirty	DOC6: I decide how others can access the work I do for this class
Dirty	DOC7: When I share my work with other students, I know how it will be used
~~Clean~~	~~COC1: I control the work I submit for assignments in this class~~
Clean	COC2: I determine who can use the work I do for this class
Clean	COC3: I regulate who can access the work I do for this class
Clean	COC4: I determine when others can access the work I do for this class
Clean	COC5: I control where others can use the work I do for this class
Clean	COC6: I decide how others can access the work I do for this class
Note: questions answered using six-point scale from strongly disagree to strongly agree.	

Knowledge: developing an intimate understanding and knowledge of an object gives rise to a sense of possession. The object of interest in this study is the assignment (the ideas, the work that demonstrates those ideas and the file itself). During the process of working through an assignment a student acquires information about the tasks and steps that constitute a solution.

This study is asking a series of questions to determine if the student (the person answering the survey) has an intimate knowledge of their assignments (the work, the file, and answers) they submit for this class.

Table C5. Psychological Ownership: Knowledge

Construct	Item
~~Dirty~~	~~DOK1: I know what goes into my work very well~~
Dirty	DOK2: I am intimately familiar with the process used to perform work tasks in my assignment files
Dirty	DOK3: I took every opportunity to oversee how things operated in my assignment work
Dirty	DOK4: I recognize the work on each page of my assignment files
Clean	COK1: I know what goes into my assignment work
Clean	COK2: I am very familiar with the tasks in my assignment files
Clean	COK3: I understand the solutions in my assignment files
Clean	COK4: I recognize the work on each page of my assignment files
Note: questions answered using six-point scale from strongly disagree to strongly agree.	

Self-identity: we own our labor and often feel we own the objects we create, shape, and produce. In the study, we focused on class assignments as the object of interest (the ideas, the work that demonstrates those ideas, and the file itself). By pouring our effort, energy, time, and perseverance into an assignment, we develop feelings that it represents us.

In the study, we asked a series of questions to determine if the students (the individuals who answered the survey) felt they invest themselves into the assignments (the work, the file, and the answers) they submit in this class.

Table C6. Psychological Ownership: Self Identity

Construct	Item
Dirty	DOSI1: The files I turn in for assignments in this class is MY work
Dirty	DOSI2: The work in files I turn in represents me
Dirty	DOSI3: The work in the files I submit in this class accurately defines my understanding of the assignment
Dirty	DOSI4: I feel a high degree of personal ownership for the assignment files I turn in for this class
~~Dirty~~	~~DOSI5: It is hard for me to think about the work in the files I turn in as MINE~~
~~Dirty~~	~~DOSI6: When classmates see my work in this class, they see my capabilities~~
~~Dirty~~	~~DOSI7: When others analyze my work in this class, they can accurately evaluate ME~~
Dirty	DOSI8: My assignment work in this class reflects what I can do
~~Dirty~~	~~DOSI9: If someone were to turn in my work as their own, it diminishes ME~~
~~Clean~~	~~COSI1: The assignment files I submit reflect my effort~~
Clean	COSI2: The assignment work in files I turn in represents ME
~~Clean~~	~~COSI3: When others analyze my work in this class, they can accurately evaluate ME~~
Clean	COSI4: My assignment work in this class reflects what I can do
Clean	COSI5: I can identify with my assignment work in this class. They are my creation.
Clean	COSI6: I personally invested a lot in the assignments for this class
Clean	COSI7: When I think about it, I see part of myself in my class assignments
Note: questions answered using six-point scale from strongly disagree to strongly agree.	

About the Authors

Kurt Schmitz is a Clinical Associate Professor at the J. Mack Robinson College of Business, Georgia State University. He holds a BS in Business Administration from Samford University, an MS in Computer Science from Rensselaer Polytecnic University, and a PhD in Information Systems from the University of Texas in Arlington. He is certified by the Project Management Institute as a program management professional (PgMP) with extensive industry experience having served in multiple global IT leadership roles in large enterprises including industrial automation in the 1980s, networking and eCommerce in the 1990s, and life sciences in the 2000s. His research interests involve adaptation and change in the area of IT project management and the IT technologies produced by those projects.

Veda Storey is the Tull Professor of Computer Information Systems and professor of computer science at the J. Mack Robinson College of Business, Georgia State University. Her research interests are in intelligent information systems, data management, conceptual modeling, and design science research. She is particularly interested in the assessment of the impact of new technologies on business and society from a data management perspective. She is a member of the steering committee of the International Conference of Conceptual Modeling.