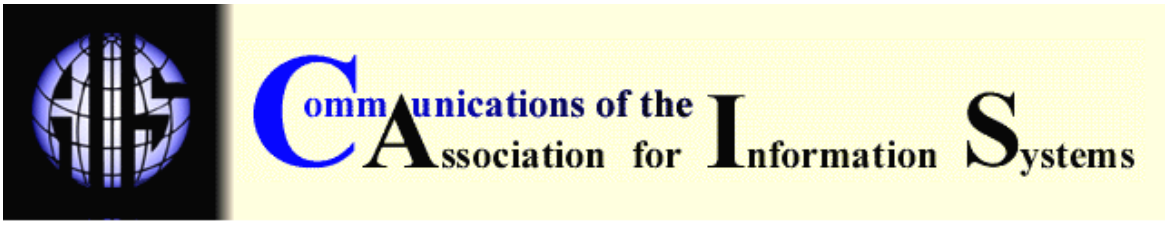


VALIDATION GUIDELINES FOR IS POSITIVIST RESEARCH

Detmar Straub
Georgia State University
dstraub@gsu.edu

Marie-Claude Boudreau
University of Georgia

David Gefen
Drexel University

ABSTRACT

The issue of whether IS positivist researchers were validating their instruments sufficiently was initially raised fifteen years ago. Rigor in IS research is still one of the critical scientific issues facing the field. Without solid validation of the instruments that are used to gather data on which findings and interpretations are based, the very scientific basis of the profession is threatened.

This study builds on four prior retrospectives of IS research that conclude that IS positivist researchers continue to face major barriers in instrument, statistical, and other forms of validation. It goes beyond these studies by offering analyses of the state-of-the-art of research validities and deriving specific heuristics for research practice in the validities. Some of these heuristics will, no doubt, be controversial. But we believe that it is time for the IS academic profession to bring such issues into the open for community debate. This article is a first step in that direction.

Based on our interpretation of the importance of a long list of validities, this paper suggests heuristics for reinvigorating the quest for validation in IS research via content/construct validity, reliability, manipulation validity, and statistical conclusion validity. New guidelines for validation and new research directions are offered.

Keywords: IS research methods; rigor; measurement; psychometrics; validation; reliability; content validity; construct validity; convergent validity; discriminant validity; nomological validity; predictive validity; concurrent validity; unidimensional reliability; factorial validity; manipulation validity; statistical conclusion validity; formative and reflective measures; quantitative, positivist research; heuristics; guidelines; structural equation modeling; LISREL; PLS.

I. INTRODUCTION

Fifteen years ago, Straub [1989] raised the issue of whether IS positivist researchers were sufficiently validating their instruments. Information systems (IS) research is a dynamic and ever changing field and since then the IS profession was exposed to many opportunities and challenges. E-Commerce rose and, at least the term “e-Commerce,” fell in prominence. Other management trends passed through their cycles. Our professional society, the Association of

Information Systems (AIS), was formed and amalgamated with the preeminent research conference in the field, the International Conference on Information Systems, and formed an alliance with one of the premier journals in the field, the *MIS Quarterly*.

But have such momentous events in the life of the profession also been reflected in dramatic improvements in scientific practices, especially those related to validation of our research instruments? A brief history of the validation phenomenon should answer this question, in part, at least. In 1989, Straub's call for new efforts to validate IS research instruments was based on a straight-forward survey of the use of various techniques for gathering empirical data. He found that only 17% of the articles in three widely referenced IS journals over the previous 3 years reported reliability of their scales, only 13% validated their constructs, while a scant 19% used either a pretest or a pilot test. The argument for validation of instruments was based on the prior and primary need to validate instruments before such other crucial validities as internal validity and statistical conclusion validity are considered. Three follow-up studies by Boudreau et al. [2004], Boudreau et al. [2001], and Gefen et al. [2000] suggest that the field is moving slowly but steadily toward more rigorous validation in positivist work. These studies also found that nearly all forms of instrument validation were still in the minority of published articles in *MIS Quarterly*, *Management Science, Information Systems Research, Journal of Management Information Systems*, and *Information & Management*.

It is important to note at the outset that we are not maintaining in any way that positivist work is superior (or inferior) to post-modern approaches. We simply do not taking any stand on that issue. The paper is addressed at positivist, quantitative researchers who already accept the epistemological stance that their line of inquiry is useful and defensible. Realizing that this approach characterizes the work of many North American academics, we imply absolutely nothing about the quality of the work in other parts of the world that may or may not adopt the quantitative, positivist intellectual position. Stated in absolute terms, this epistemological stance is that the world of phenomena involves an objective reality that can be measured and that relationships between entities in this world can be captured in data that is reasonably representative and accurate. Since the entities of significance can be present in data about them, the causal linkages between entities can also be assessed. This simple assertion presents the most extreme version of this line of inquiry. It is perhaps fitting to indicate that many modern positivist researchers are willing to accept the possibility that many of the entities they articulate are social constructions. However, the "permanent" presence of these constructs in the real world allows them to be evaluated along the same lines as harder and less demonstrably subjective realities. In short, many contemporary positivist researchers are willing to consider constructs as a "fuzzy set" rather than as the near perfect surrogate of an objective reality.

These concessions do not in any way diminish the strength of belief of many positivist researchers that we are able to capture approximations of real world entities, many of which are intellectual (or social) constructions to be sure. Capturing these entities is a process that can be fraught with difficulties. One of the most tenacious of these is the inability of the IS community to know whether the measures being selected and used by the researchers are valid. The concept is that straight-forward.

- Valid measures represent the essence or content upon which the entity or construct is focused.
- They are unitary.
- They are not easily confused with other constructs.
- They predict well.
- If they are supposed to manipulate the experience of subjects, they do so.

The current study builds on analyses of IS research since the turn of the millennium, namely, Gefen et al. [2000], Boudreau et al. [2001], and Boudreau et al. [2004], which, in brief, conclude

that IS positivist researchers still have major barriers to overcome in instrument, statistical, and other forms of validation. The present study goes far beyond Straub [1989], Gefen et al. [2000], Boudreau et al. [2001], and Boudreau et al. [2004] by extending these prior articles through discussion of other critical validities such as:

- Nomological validity
- Split-half reliability
- Test-retest reliability
- Alternate forms of reliability
- Inter-rater reliability
- Unidimensional reliability
- Predictive validity
- Manipulation validity

The main contribution of this study is to offer research heuristics for reinvigorating the quest for validation via

- content validity,
- construct validity,
- reliability,
- manipulation validity, and
- statistical conclusion validity[1].

These heuristics are based on a thorough analysis of where the field stands with respect to all key instrument validities. Some of these heuristics will, no doubt, be controversial. We believe that it is time for the IS academic profession to bring such issues into the open for community debate, and this article is an initial step in that direction.

To build our case, it is necessary first to discuss each validity at some length (Section II). Specifically, content validity, construct validity, predictive validity, reliability, manipulation validity, and statistical conclusion validity are presented. Discussion of these validities serves as a reference point for proposing specific heuristics in each validation category (Section III). To provide extra guidance, each heuristic is qualified as being mandatory, highly recommended, or optional**.** Then, to demonstrate that these heuristics are attainable, an example of how instruments can be developed is presented in Section IV. The final section offers concluding remarks.

II. REVIEW AND REASSESSMENT OF VALIDATION PRINCIPLES

Viewed from the perspective of the long history of the philosophy of science, validation of positivist research instruments is simply a late 20th century effort of the academic disciplines to understand the basic principles of the scientific method for discovering truth [Nunnally, 1978]. Assuming that nature is to some extent objectively verifiable, the underlying truths of nature are thought to be revealed slowly and incrementally, with the exception of occasional scientific revolutions of thought [Kuhn, 1970]. This process of "normal" positivist science is also believed to result from successful paradigms that invoke theories, in which causally-linked intellectual

[1] A glossary of the terms used in this article is presented at the end of the paper. This glossary includes validation concepts and techniques, and heuristics.

constructs represent underlying natural [Kuhn, 1970], artificial [Simon, 1981], or social phenomena [Blalock, 1969]. "Normal" science also includes a consideration of methods favored by given disciplines and deemed to be valid for the discovery of truth [Scandura and Williams, 2000]. In positivist science, the need to ensure that the data being gathered is as objective as possible and a relatively accurate representation of the underlying phenomenon is paramount.

Social science or behavioral research, which describes a significant proportion of all IS research, can be more or less rigorously conceived or executed. The rigor of the research design is often characterized by the extent to which the data being gathered is an accurate representation of latent constructs that may draw on numerous sources and kinds of data, and relevant to the theory that the researcher is attempting to build or test [Coombs, 1976]. Latent constructs are latent in the sense that they are not directly observable. In short, there are no immediate or obvious measures that the scientific community would agree on that capture the essence of the construct. The construct itself can be viewed as a social construction, represented by a set of intellectually-derived measures that are not self-evident or inherently "true" measures. Measures are, therefore, indirect; they are surrogates, to a greater or lesser extent, of the underlying research construct.

While these points express fundamental validation principles, they do not indicate specifically how a researcher attempting to use valid scientific methods should proceed at a pragmatic, concrete level. This paper attempts to address this issue by setting forth specific guidelines, i.e., heuristics, for validation which are based on both intellectual soundness and best of breed IS research practice.

How would one know which validation principles make sense, both on an individual basis and on the basis of the field as a whole? The social sciences tend to develop validation principles concurrent with the pursuit of research. Hence, practice and terminology vary widely. Ironically, though, this question cannot be answered simply because scientific methods and techniques cannot themselves be used to validate the principles upon which they are based. Scientific principles for practice are only accepted as received wisdom by a field or profession through philosophical disputation [Nunnally, 1978]. Over time, they become accepted norms of conduct by the community of practice.

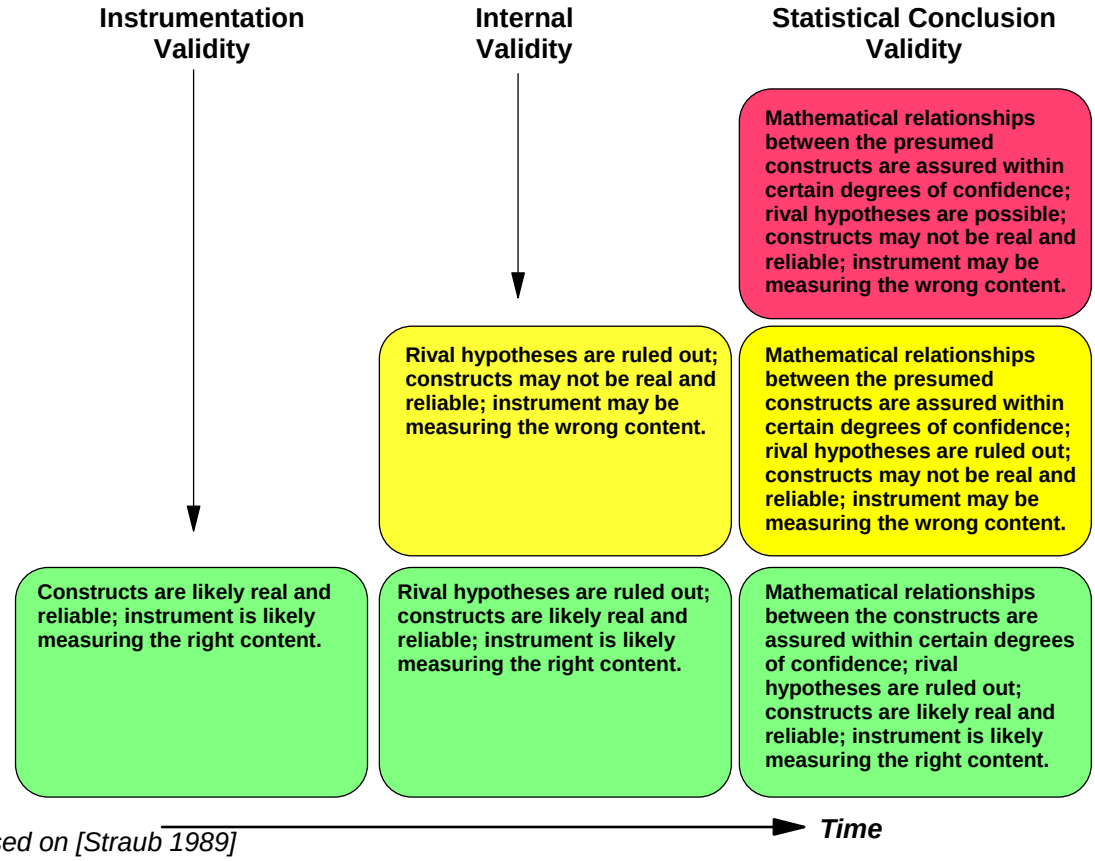
Articulation of validation principles and acceptance of validation ideas by the IS field depend strictly on calls to authority and the persuasiveness of the ideas themselves. There are no established scientific standards against which to test or evaluate them.

VALIDITY AND VALIDITY TOUCHSTONES

The purpose of validation is to give researchers, their peers, and society as a whole a high degree of confidence that positivist methods being selected are useful in the quest for scientific truth [Nunnally, 1978]. A number of validities are discussed by Cook and Campbell [1979]. These terms and the respective terms used for these validities are:

validation of data gathering	instrument/instrumentation validity
ruling out rival hypotheses	internal validity
statistical inference	statistical conclusion validity
generalizability	external validity.

Straub [1989] argues for an order of precedence in which these validities should be considered. The basic case is that instrument validation is both a *prior and primary validation* for IS empirical research. In other words, if validation of one's instrumentation is not present *or* does not precede internal validity and statistical conclusion validity, then all other scientific conclusions are thrown into doubt (cf. also Andrews [1984]). These "Validity Touchstones" and the consequences if they are considered or ignored are shown in Figure 1.



Based on [Straub 1989]
Green is preferred path; yellow is cautionary; red is least desirable path

Figure 1. Validity Touchstones

To articulate the specific validation principles being referred to, we next briefly review and reassess the validation principles discussed in Straub [1989]. This discussion provides completeness, adds to the previous work, and presents new thoughts on IS instrument validities in the context of heuristics for practice. Based on this review and on building the cases for heuristics, findings from empirical studies of the use of validities in IS will be more meaningful. Validities to be reviewed along with heuristics and typical techniques are discussed with exemplars in IS research in Table 1.

CONTENT VALIDITY

Content validity is an issue of representation. The essential question posed by this validity is:

> Does the instrumentation (e.g., questionnaire items) pull in a representative manner from all of the ways that could be used to measure the content of a given construct [Cronbach, 1971, Kerlinger, 1964] ?

Table 1. Validities, Heuristics, and Examples in IS Research

Validity Component	Heuristics/Techniques	Comments/Pros and Cons	Examples in IS Research
Content Validity	Literature review; expert panels or judges; content validity ratios [Lawshe, 1975]; Q-sorting	Infrequent in IS research.	[Smith et al., 1996]; [Lewis et al., 1995]; [Storey et al., 2000]
Construct Validity Discriminant validity (sometimes erroneously called divergent validity)	MTMM; PCA; CFA as used in SEM; PLS AVE analysis; Q-sorting	MTMM rare in IS research; no well accepted statistical thresholds for MTMM, but without at least a two method comparison, other techniques do not account as well for common methods bias [for an opposing argument, see Bagozzi et al. [1991].	[Igbaria and Baroudi, 1993]; [Venkatraman and Ramanujam, 1987]; [Straub, 1990]
Convergent validity	MTMM; PCA; CFA as used in SEM; Q-sorting	Rare in IS research. No well accepted statistical thresholds for MTMM, but without at least a two method comparison other techniques do not account as well for common methods bias.	[Igbaria and Baroudi, 1993]; [Venkatraman and Ramanujam, 1987]; [Straub, 1990]; [Gefen, 2000]
Factorial validity	PCA; CFA as used in SEM	Favored technique in IS research; assesses discriminant and convergent validity; common methods bias remains a threat to validity without at least a two method comparison.	[Brock and Sulsky, 1994]; [Adams et al., 1992]; [Doll and Torkzadeh, 1988]; [Barki and Hartwick, 1994]
Nomological validity	Judgmental comparison with previous nomological (theoretical) networks; patterns of correlations; regression; SEM	Infrequent, likely because of the lack of widely-accepted theory bases in IS.	[Igbaria and Baroudi, 1993]; [Straub et al., 1995]; [Pitt et al., 1995]; [Smith et al., 1996]
Predictive validity (a.k.a. concurrent or post-diction validity)	Correlations; Z-scores; discriminant analysis; regression; SEM	Useful, especially when there is a practical value to the prediction; used little in the past, but becoming more frequent in IS research.	[Szajna, 1994]; [Pitt et al., 1995]; [Van Dyke et al., 1997]; [Collopy et al., 1994]; [Smith et al., 1996]
Common methods bias / method halo	MTMM, CFA through LISREL	Notably rare in IS and related research, especially when data is collected via surveys.	There is an excellent example in the psychological literature: [Marsh and Hocevar, 1988]; see, however, Woszczynski and Whitman [2004]
Reliability Internal consistency	Cronbach's α; correlations; SEM composite consistency estimates	α assumes that scores for all items have the same range and meaning; if not true, adjustments can be made in the statistics; also, nonparametric correlations can be plugged into the formulation.	[Grover et al., 1996]; [Sethi and King, 1994]

Validity Component	Heuristics/Techniques	Comments/Pros and Cons	Examples in IS Research
Split half	Cronbach's α; correlations	Different results may be obtained depending on how one splits the sample; if enough different splits are made, the results approximate internal consistency Cronbach's α.	[McLean et al., 1996]
Test-retest	Cronbach's α; correlations; SEM estimates	Comparisons across time of an instrument.	[Hendrickson et al., 1993]; [Torkzadeh and Doll, 1994]
Alternative or equivalent forms	Cronbach's α; correlations; SEM estimates	Comparisons across time and forms of an instrument.	[Straub, 1989]
Inter-rater reliability	Percentages; correlations; Cohen's Kappa	Transformation of correlations suggested.	[Massetti, 1996]; [Lim et al., 1997]; [Boudreau et al., 2001]
Unidimensional reliability	SEM, as performed in LISREL	Novel, sophisticated technique for assessing reliability .	[Segars, 1997]; [Gefen, 2000]; [Gefen, 2003]
Manipulation Validity (a.k.a. manipulation checks)	Percentages; t-tests; regression; discriminant analysis	No standard procedures are agreed upon; practice varies significantly.	[Keil et al., 1995]; [Straub and Karahanna, 1998]

As Figure 2 shows, researchers have many choices in creating means of measuring a construct. Did they choose wisely so that the measures they use capture the essence of the construct? They could, of course, err on the side of inclusion or exclusion. If they include measures that do not represent the construct well, measurement error results. If they omit measures, the error is one of exclusion.

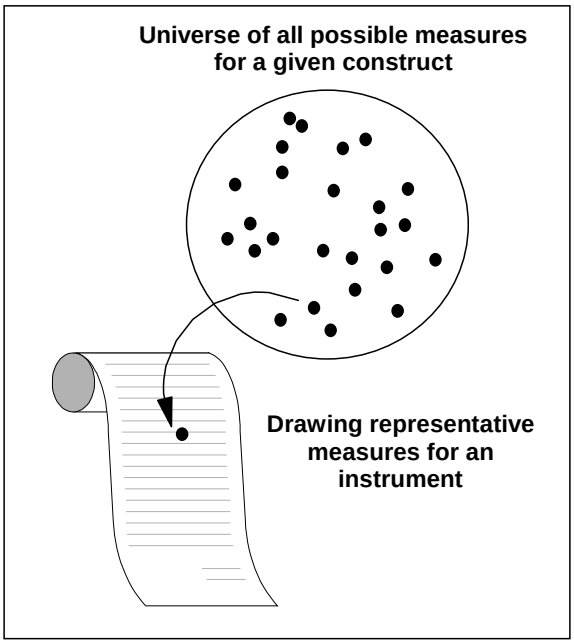


Figure 2. A Pictorial Model of Content Validity

Whereas many psychometricians [Barrett, 1980-81, Barrett, 1981, Nunnally, 1978] indicate that content validity is a valuable, albeit complex tool for verifying one's instrumentation, others argue

that it is a concept that cannot be validated because it deals essentially with an essentially unknowable sampling issue and not an instrument evaluation issue [Guion, 1977].

As discussed in Straub [1989], content validity is established through literature reviews and expert judges or panels. Several rounds of pretesting the instrument with different groups of experts is highly advisable. Empirical assessment of this validity is generally not required, although Lawshe [1975] provides a procedure and statistic for testing the level of validity.

Clearly, if there is such a thing, content validity is not easy to assess. With what degree of certainty can a researcher know that he or she is drawing representatively from the "content universe" ([Cronbach, 1971], p. 455) of all possible content? Even if experts, panels of judges, and/or field interviews with key informants are used, as recommended [Cronbach, 1971, Straub, 1989], it is not guaranteed that the instrument items are randomly drawn because the universe of the items itself is indeterminate [Cronbach, 1971, Lawther, 1986, Nunnally, 1978, Straub, 1989]. The most commonly employed evaluation of this validity is judgmental and is highly subjective.[2] Moreover, it may well be, as Guion [1977] asserts, that content validity is in essence merely content sampling and, ultimately, an evaluation of construct validity.[3] Carrier et al. [1990] present evidence that content validity is significantly correlated with predictive validity, so it may, indeed, be the case that content validity is not a validity in its own right.

In their 2001 assessment of the practice of instrument validation in IS, Boudreau et al. [2001][4] indicate that only 23% of the articles they sampled examined content validity. As for pretesting, the "preliminary trial of some or all aspects of an instrument" [Alreck and Settle, 1995], a technique which often leads to content validity, Boudreau et al. [2001] did not find this process widespread, in that only 26% of their sampled articles used such a technique.

Vignette #1: Example of Content Validity

A good example of content validation can be found in Lewis et al.'s work [1995]. These authors validated the content of their information resource management (IRM) instrument via Lawsche's quantitative approach [1975]. In this research, panelists scored a set of items derived from a literature review of the IRM concept, using the scale "1=Not relevant, 2=Important (but not essential), and 3=Essential." From these data, a content validity ratio (CVR) was computed for each item using Lawsche's formulation [1975]. Based on a table in Lawsche [1975], the CVR for each item was evaluated for statistical significance (.05 alpha level), significance being interpreted to mean that more than 50% of the panelists rate the item as either essential or important.

Heuristics for Content Validity

Having valid content is desirable in instruments for assuring that constructs are drawn from the theoretical essence of what they propose to measure. In spite of detractors, many seem to resonate with the idea of content validity. Therefore, at this point in the history of the positivist sciences, lacking clear consensus on the methods and means of determining content validity, we would argue that it is a highly recommended, but not mandatory practice for IS researchers.

---

[2] On the other hand, see Lawshe [1975], who proposes quantitative measures for this validity.
[3] Rogers [1995] also considers content validity, along with criterion-related validity, as subtypes of construct validity; for him, construct validity has become "the whole of validity."
[4] Boudreau et al. [2001] coded positivist, quantitative research articles for use of validation techniques. They examined five major journals over a three year period from 1997 to 1999.

CONSTRUCT VALIDITY

Construct validity is an issue of operationalization or measurement *between* constructs. The concern is that instrument items selected for a given construct are, considered together and compared to other latent constructs, a reasonable operationalization of the construct [Cronbach and Meehl, 1955]. Validation is not focused on the substance of the items, other than, perhaps, its meaningfulness within its usual theoretical setting [Bagozzi, 1980]. In general, substance and straightforward definitions of the construct are matters for content validity.

As illustrated in Figure 3, construct validity raises the basic question of whether the measures chosen by the researcher "fit" together in such as way so as to capture the essence of the construct. Whether the items are formative or reflective, other scientists want to be assured that, say, yellow, blue, or red measures are most closely associated with their respective yellow, blue, or red latent constructs. If, for instance, blue items, in the presence of other variables like the yellow construct, load on or are strongly associated with the blue construct, then we would say that they "converge" on this construct (convergent validity).[5] If theoretically unrelated measures and constructs are considered alongside this variable, such as with latent construct C, then there should be little or no crossloading on constructs A or B. In other words, the measures should "discriminate" among constructs (discriminant validity). Note that constructs A and B are posited to be linked, and, therefore, the test for discriminating between theoretically connected independent variables (IVs) and dependent variables (DVs) is a robust one indeed, and it may not always work out. In general, therefore, it is best not to mix IVs and DVs in factoring. The most reasonable test for whether the links are similar to those found in past literature (known as a "nomological network" of theoretical linkages) looks at path significance. If the path, indicated by the red arrow in Figure 3, is significant, then we can say that construct validity has been established through nomological validity as well.

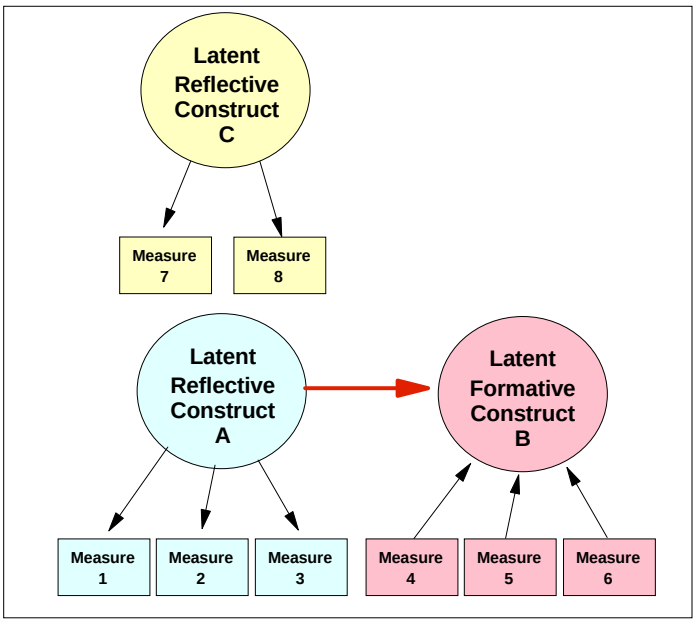


Figure 3. Pictorial Model of Construct Validity

[5] It is useful to compare an independent variable against other independent variables as well. Likewise for dependent variables. Mixing independent variables and dependent variables is often done, but not a practice we recommend.

It should be noted that nomological validity resembles hypothesis testing in that it focuses on the paths. The stress in this form of validity, though, is slightly different in that it focuses on likeness or lack of similarity to strength of construct linkages in the past literature. Comparisons with, for example, the explained variances in prior work would be appropriate for an analysis of nomological validity.

Differences from Internal Validity and Forms of Construct Validity

Construct validity differs from internal validity in that it focuses on the measurement of individual constructs while internal validity focuses on alternative explanations of the strength of links between constructs [Straub, 1989]. Internal validity can be easily mistaken for construct validity [cf. [Smith et al., 1996] as a case in point where both validities are stated to be testing the relationships between constructs], but their focus is really quite different. In establishing internal validity, the researcher is trying to rule out alternative explanations of the dependent variable(s). In establishing construct validity, the researcher is trying to rule out the possibility that constructs, which are artificial, intellectual constructions unobservable in nature, are being captured by the choice of measurement instrumentation. Nomological validity, which is one form of construct validity, does test strength of relationships between constructs but only to examine whether the constructs behave as they have in the past, that is, within the nomological or theoretical network that the researchers have defined.

Besides nomological validity, discriminant, convergent, and factorial validity are all considered to be forms of or variations on construct validity. Moreover, criterion-related validity and its sub-types, predictive and concurrent validity [Cronbach, 1990, Rogers, 1995] are also considered to be constituents of construct validity.[6] In Boudreau et al., [2001], 37% of the articles sampled by the authors established construct validity based on one (or many) of its aforementioned constituents, which are described next.

Discriminant Validity

One test of the existence of a construct is that the measurement items posited to reflect (i.e., "make up") that construct differ from those that are not believed to make up the construct.[7] Campbell and Fiske's multitrait-multimethod analysis (MTMM) [1959] can be helpful in understanding the basic concept of discriminant validity and is one mechanism for testing it. In their seminal article, Campbell and Fiske [1959] argue that choice of method (common methods bias) is a primary threat to construct validity in that study participants will, under inapt circumstances, tend to respond in certain patterns if the instrumentation, unwittingly, encourages such responses.

As an example, consider typical empirical circumstances underlying tests of TAM. Most researchers use a single instrument to query respondents or subjects about acceptance of a particular technology or application. Thus, questions about perceived usefulness and ease of use are followed by questions about intention to use a system or its actual usage. This means of testing TAM involves inherent common methods biasing because all measures are self-reported and undoubtedly tied together in the minds of the respondents. Consider whether subjects can truly separate a question about use from a question about how easy to use or how useful a system is. Cognitive dissonance theory would suggest that respondents who felt that the system

[6] It should be noted that some conceptualize predictive validity as separate and distinct from construct validity (e.g., [Bagozzi, 1980, Campbell, 1960, Cronbach, 1990]). Other methodologists, however, believe that it may be an aspect of construct validity in that successful predictions of links of constructs to variables outside a theoretical domain also validates, in a sense, the robustness of the constructs [Mumford and Stokes, 1992].

[7] See the discussion in Gefen et al. [2000] explaining the distinction between reflective and formative variables.

was useful would feel the need to answer in the affirmative that they planned to use it, and vice versa. To do otherwise, would require them to deal with an uncomfortable cognitive dissonance.

To test formally an instrument for common methods bias, two methods (i.e., instruments or data gathering-coding methods) are required [Straub, 1989]. These methods should be "maximally different" ([Campbell and Fiske, 1959], p. 83) so that the distinctions in the underlying true scores attributable to method are revealed. Measures (termed "traits" in MTMM) show discriminant validity when the correlation of the same trait and varying methods is:

- significantly different from zero and
- higher than that trait and different traits using both the same and different methods.

Tests of discriminant validity in IS research typically do not use MTMM, perhaps because its rules of thumb are ambiguous [Alwin, 1973-74] and it is labor-intensive, requiring two very different methods of gathering all data. In certain cases, it may be the only approach available to test discriminant validity, though. When measures are formative rather than reflective [Diamantopoulos and Winklhofer, 2001, Fornell and Larcker, 1981, Gefen et al., 2000], for instance, the measures "causing" the latent construct may lack high inter-correlations and assume different weights, as in a regression with multiple independent variables. Methods used to test validity of formative measures rely on principles articulated in the original Campbell and Fiske MTMM technique [1959]. Podsakoff et al. [2003] present models for testing common methods bias, approaches which can be useful in proving the validity of formative measures.

One method for creating a weighted, summed composite score for the "latent" construct is suggested by the mathematical formulation in Bagozzi and Fornell [1982]. These composites scores can be compared against a normalized score for each measure to be certain that items relate more strongly to their own latent construct than to other constructs. Ravichandran and Rai [2000] offer another technique for testing formative measures along this line of thinking.

Finally, Loch et al. [2003] offer an alternate discriminant validity test of formative latent variables using PLS weights. In their approach, weights from a formative PLS model of the indicators (measures) is used to derive/calculate a latent construct value for each variable. These values are then compared using a modified MTMM analysis. In the case of this particular study, latent variables were sufficiently different in posited directions to argue that the instrument is valid.

Other techniques besides these three can be used to evaluate discriminant validity. In that many of these techniques are based on variants of factor analysis, they will be discussed below under "factorial validity." But one such innovative means of verifying discriminant validity is Q-sorting [Moore and Benbasat, 1991, Segars and Grover, 1998, Storey et al., 2000, Thomas and Watson, 2002]. Q-sorting combines validation of content and construct through experts and/or key informants who group items according to their similarity. This process also eliminates (i.e., discriminates among) items that do not match posited constructs.

Features of covariance-based SEM likewise permit the assessment of discriminant validity. It is shown by comparing two models, one which constrains the item correlations to 1 and another which frees them, i.e., permits them to be estimated [Segars, 1997]. By comparing the χ^2s of the two models, it is possible to test for discriminant validity. A significant difference between the models, which is also distributed as χ^2 [Anderson and Gerbing, 1988], indicates that the posited construct items are significantly different from other construct items in the overall model. This analysis is becoming more frequent in IS research (see Gefen et al., [2000], for a running example and Gefen [2003] for a mainstream publication that assesses this analysis.)

Vignette #2: Examples of Discriminant Validity

In several examples in IS research other than Straub [1989] and Straub [1990], a formal MTMM analysis was used via two extremely different methods, namely comparisons of pencil-and-paper questionnaire responses and interview responses to the same questions. Igbaria and Baroudi [1993] developed an instrument to measure career anchors and employee's self-concepts. The nine career anchors were: technical competence; managerial; autonomy; job security; geographic security; service; pure challenge; life-style; and entrepreneurship. Arguing that they are using MTMM, they compared within-construct interitem correlations to between-construct inter-item correlations. Examination of the correlation matrix of 25 items showed that of the 300 comparisons, only 9 did not meet the 50% violation-criteria specified by Campbell and Fiske [1959]. A legitimate question to raise in this context is whether Igbaria and Baroudi [1993] are truly using multiple, maximally different methods in evaluating their instrument.

An example of a study using a different technique than MTMM to assess discriminant validity is Segars and Grover [1998]. They use Q-sorting to validate the construct and sub-constructs of strategic information systems planning (SISP). Their literature review found four dimensions of SISP with 28 associated planning objectives. A random listing of the 28 objectives in single sentence format were provided on pages separate from the 4 sub-construct dimensions of: (1) alignment, (2) analysis, (3) cooperation, and (4) improvement capabilities. Experts and key informants were asked to sort the objectives into the four dimensions. The overall percentage of correct classification was 82%, with individual items correctly classified at a rate of 90% or better being retained. Twenty-three objectives exhibited consistent meaning across the panel and were adopted as measures of their associated constructs.

Convergent Validity

Convergent validity is evidenced when items thought to reflect a construct converge, or show significant, high correlations with one another, particularly when compared to the convergence of items relevant to other constructs, irrespective of method.[8] The comparison with other constructs is one element that distinguishes convergent validity from reliability.

A classic method for testing convergent validity is MTMM analysis. As discussed at length in Campbell and Fiske [1959] and Straub [1989], this highly formal approach to validation involves numerous comparisons of correlations and correlational patterns. Percentages smaller than chance of violations of convergent and discriminant validity conditions in the matrix of trait (or item) and method correlations indicate that the methods are equally valid.

Problems with MTMM are legion. Bagozzi [1980] and Bagozzi and Phillips [1982] argue that counting violations in a correlation matrix is an arbitrary procedure that can lead to incorrect conclusions. If a researcher gathered data via more than one method, Bagozzi [1980] shows how SEM can be used to examine method versus trait variance as well as other validity properties of the entire MTMM matrix of correlations. SEM indeed permits the assessment of convergent validity: when the ratio of factor loadings to their respective standard errors is significant, then convergent validity is demonstrated [Segars, 1997]. The most thorough analysis of tests for methods bias within the context of convergent validity appears in Podsakoff et al. [2003].

[8] Convergent validity is important for reflective variables, but less so for formative ones. In fact, one definition of formative constructs is that the measures need not be highly correlated. Socio-economic status is measured by such items as household income and the number of children per household; these are both indicators of this status, but may not be correlated ([Jöreskog and Sörbom, 1989]). See Gefen et al. [2000] for definitions and Diamantopoulos and Winklhofer [2001] for a detailed discussion.

As with discriminant validity, MTMM analysis of convergent validity is infrequent in IS research. MTMM's requirement for gathering of data through at least two "maximally different methods" ([Campbell and Fiske, 1959], p. 83) places such a heavy burden on researchers that they may be shying away from it. In fact, no matter how much the community wishes to have valid instruments, it is possibly overmuch to ask researchers to engage in this level of validation at an early stage in a research stream. Several examples involving MTMM application are described below, but IS researchers probably only need to apply MTMM after a research stream matures [Straub, 1989]. A more exacting need is to rule out methods bias. Clearly, a pressing need is to scrutinize the entire TAM research stream for common methods bias (e.g., [Straub et al., 1995]).

Vignette #3: Examples of Convergent Validity

A case where MTMM was used to assess convergent validity was [Venkatraman and Ramanujam, 1987]. These authors are true to the letter and spirit of MTMM in gathering their data via very different sources (methods). Self-reported data are compared to archival (COMPUSTAT) firm data for three measures of business economic performance (BEP) — sales growth, profit growth, and profitability. The MTMM analysis found strong support for convergent validity and moderate support for discriminant validity. What this finding suggests is that method plays little to no role in measures of BEP, i.e., subjective managerial assessments of performance are equal in validity to objective measures.

Another MTMM analysis was reputedly performed in Davis [1989]. Comparing user responses to the technologies of Xedit and E-mail, he treats these technologies as if they were distinct and separate data gathering "methods," in the sense of Campbell and Fiske [1959]. When Campbell and Fiske [1959] speak of "maximally different" methods, however, they clearly have in mind different types of instrumentation or source of information, such as pencil-and-paper tests versus transcribed interviews, or course evaluations by peers versus evaluations by students. Different technologies are likely not different "methods." Nevertheless, Davis [1989] examines the correlations of 1800 pairs of variables, finding no MTMM violations in his tests for convergent validity (monotrait-monomethod triangle) and for discriminant validity of perceived usefulness. He found only 58 violations of 1800 (3%) for discriminant validity of perceived ease-of-use, which he interprets as acceptable.

Factorial Validity

While factorial validity was discussed briefly in Straub [1989], several points of clarification are definitely in order. Factorial validity can assess both convergent and discriminant validity, but it cannot rule out methods bias when the researcher uses only one method,[9] which is by far the most frequent practice in IS research. Moreover, if two or more methods are used to assess the instrument in question,[10] there is evidence that MTMM is preferable to factor analytic techniques [Millsap, 1990]. Conversely, when only one method can be used in conducting the research, factorial techniques are likely more desirable than MTMM [Venkatraman and Ramanujam, 1987].[11]

Nevertheless, construct validity, specifically convergent and discriminant validity, can be examined using factor analytic techniques such as common factor analysis, PCA, as well as confirmatory factor analysis in SEM, such as LISREL and PLS. Convergent and discriminant

[9] It should be noted that whether methods bias is a significant problem in organizational research is an ongoing debate [Spector, 1987, Woszczynski and Whitman, 2004].

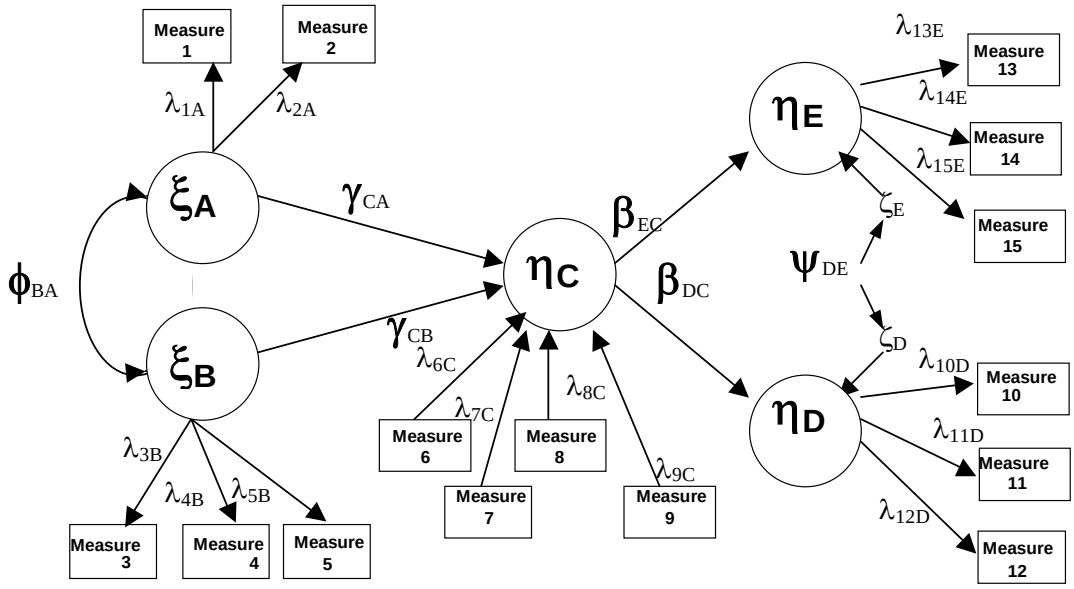
[10] Since validation is "symmetrical and egalitarian" ([Campbell, 1960], p. 548), all data gathering/coding methods are actually being validated when an MTMM is used in assessment. Nevertheless, only one of these methods may be intended for a major data-gathering effort, as in Straub [1989].

[11] Several authors argue that MTMM involves limitations [Alwin, 1973-74], some of which have been solved by CFA and structural equation modeling [Bagozzi and Phillips, 1982].

validity are established by examining the factor loadings to ensure that, once cross-loading items are dropped, items load cleanly on constructs (factors) upon which they are posited to load and do not cross-load on constructs upon which they should not load.[12]

The point of factorial validity is to examine how variables in each distinct causal stage of the theoretical network behave. What is not important is how measures may or may not cross-load among these stages. In testing the construct validity of constructs A, B, C, D and E in Figure 4, for example, it is to be expected that measures for construct A might correlate highly with those of construct C, in that there is a posited causal link between the constructs. It is even conceivable that some measures[13] in construct A will correlate more highly with those in construct C than with other measures in its own construct, construct A. Therefore, it is of primary interest to test construct A against construct B, which is another independent variable in the same causal stage explaining construct C.

Having said that, it is important to recognize that covariance-based SEM takes into account all covariances, including those not explicitly specified in the model. Consequently, if the items in construct A are highly correlated with those in construct D, neglecting to specify explicitly that construct A is correlated with construct D will result in unacceptably low fit-indices. (Gefen et al., [2000], discuss and compare covariance-based SEM, PLS, and linear regression).



***Exogenous** Latent Variables A and B* ***Endogenous** Latent Variables C, D, and E*

Adapted from Gefen et al. [2000]

Figure 4. Generic Theoretical Network with Constructs and Measures

[12] How to handle items that do not load properly is a matter of some debate. The issue is whether items that do not load properly should be dropped (as suggested by Churchill [1979] and by Gerbing and Anderson [1988] or not, as suggested by MacCallum and Austin [[2000]). The important point to note is that factor analysis can show the way to "clean up" the construct.

[13] If *most* of the variables across causal stages cross-loaded, then there could well be a serious problem with common methods bias [Campbell and Fiske, 1959]. As in MTMM analysis, the homotrait, homomethod correlations should always be highest in the matrix of methods/traits. But cross-loading of *some* of the variables does not seriously threaten the validity of the instrument [Campbell and Fiske, 1959].

The point of factorial validity is to examine the constructs independent of the theoretical connections. When PCA is used, in this case as an exploratory factor analysis technique, researchers can simply test the groups of variables separately. In Figure 4, measures 1 through 5 (constructs A & B) should be run in a separate PCA from the measures for construct C. Assuming that there are measures in the instrument other than those in this theoretical model, Construct C should be run separately from D and E, as well.[14] The validation question the researcher is asking is whether the pattern of factor loadings corresponds with the *a priori* structure of latent constructs in each stage in the causal chain. The theoretical question is whether the constructs are related. Loadings across what are traditionally known as independent and dependent variables are, therefore, not relevant to the issue of construct validity and such tests may/should be avoided in PCA.

SEM, on the other hand, facilitates examining factorial validity through a Confirmatory Factor Analysis (CFA). That is, examining the "correctness" of the measurement model (specifying for each item its corresponding construct) that the researcher specified. In the case of covariance-based SEM, such as LISREL, CFA is first run where the researcher explicitly specifies the measurement model and runs the SEM in CFA mode. The fit statistics of this CFA provide a good indication of the extent to which the measurement model accounts for the covariance in the data. If the fit statistics are below the accepted thresholds, the research model is not supported by the data. Covariance-based SEM techniques also allow a more detailed examination of the measurement model by comparing the χ^2 statistic of the proposed measurement model to alternative ones [Bagozzi, 1980, Gefen et al., 2000, Segars, 1997]. In the case of PLS, SEM facilitates the examination of factorial validity by allowing the researcher to specify *a-priori* which items should load on which construct and then examining the correlations and the Average Variance Extracted (AVE). Factorial validity is established when each item correlates with a much higher correlation coefficient on its proposed construct than on other constructs and when the square root of each construct's AVE is notably larger than its correlation with other constructs [Chin, 1998a, Gefen et al., 2000, Karahanna et al., 1999]. Gefen et al. [2000] offer a detailed discussion.[15]

Vignette #4: Example of Factorial Validity

As an example of an empirical test using factorial validity, [Gefen et al., 2000] is worth examining because of its factorial comparisons across LISREL, PLS, and traditional factorial validity approaches like PCA. In this case, the TAM measures are validated for use in a free simulation experiment in an e-commerce setting. Principal components factor analysis verified the construct validity of the instrument for the regression tests of the posited TAM linkages. In PLS and LISREL, the item loadings on the latent construct were sufficiently high and significant to indicate acceptable measurement properties.

Nomological Validity

Although not discussed in Straub [1989], nomological validity is a form of construct validity that is beginning to be seen more frequently for assessing construct validity.[16] As described in

[14] Control variables are often useful in running such tests of discriminant validity.

[15] It should be noted that many researchers use factorial validity to test convergent and discriminant validity. Factors that load cleanly together (and do not cross-load) are said to be evidence of "convergent" validity. Those that do not cross-load are evidence of "discriminant" validity.

[16] Westland [2004] argues that nomological validity was abandoned in psychology as a result of MTMM, but there are numerous examples of this form of validity being used in recent years in this literature [Dholakia and Bagozzi, 2003, Netemeyer et al., 2002, Nielsen et al., 2000, Webster and Compeau, 1996]. Moreover, it has been advocated and practiced in the business disciplines, notably Bagozzi [1980]; Netemeyer [1991]; McKnight [2002b]; Chin [1997]; Pitt [1995]; and Devaraj [2002]. We list other examples as well in the IS

Cronbach and Meehl [1955], Cronbach [1971], and Bagozzi [1980], nomological validity is construct validity that devolves from the very existence of a well developed theoretical research stream (also called a nomological "network"). If theoretically-derived constructs have been measured with validated instruments and tested against a variety of persons, settings, times, and, in the case of IS research, technologies, then the argument that the constructs themselves are valid becomes more compelling. This argument is even stronger when researchers choose different methods for measuring their constructs [Sussmann and Robertson, 1986].

Assume that one researcher uses a structured interview script to gather data on a construct. Suppose that another researcher in another setting uses a questionnaire instrument. Clearly, the method of measurement is very, even maximally different. Yet, if both studies find significant linkages between the constructs using different measures, then both may be said to be "nomologically" valid. According to Campbell [1960], validation always works in both directions: it is "symmetrical and egalitarian" (p. 548).

The same robustness would be demonstrated if a researcher using a variant form of construct measurement found similar significance as studies that had used the same validated instrument. A good example of this would be Straub et al. [1995] who use a variant of Davis' TAM instrument for self-reported measures of perceived usefulness, perceived ease of use, and perceived systems usage. In spite of using variants of Davis' instrument items, the strength of the theoretical links in this study were similar to those of other works in this stream. The inference that can be made from this similarity of findings is that, in testing the robustness of the instrumentation, the new study helps to further establish the nomological validity of the constructs.

Vignette #5: Examples of Nomological Validity

Igbaria and Baroudi [1993] examined nomological validity in their instrument development of an IS career orientations measurement instrument. They found that six of nine correlations corresponded with those found between and among a variety of theoretical constructs in the literature. Thus, this helps to establish the nomological validity of their instrument.

A more recent study by McKnight et al. [2002a] examines the psychometric properties of a trust instrument. To prove that trust is a multi-dimensional concept, these authors test the internal nomological validity of relationships among the trust sub-constructs. For external nomological validity, they look at relationships between the trust constructs and three other e-commerce constructs -- web experience, personal innovativeness, and web site quality.

Ruling Out Common Methods Bias / Method Halo

As explained above in passages dealing with MTMM, common methods bias, also known as "method halo" or "methods effects," may occur when data are collected via only one method [Campbell and Fiske, 1959] or via the same method but only at one point in time [Marsh and Hocevar, 1988]. Data collected in these ways likely share part of the variance that the items have in common with each other due to the data collection method rather than to:

- the hypothesized relationships between the measurement items and their respective latent variables or
- the hypothesized relationships among the latent variables. As a result of such inflated correlations, path coefficients and the degrees of explained variance may be overstated in subsequent analyses [Marsh and Hocevar, 1988].

---

literature later. This is not to say that nomological validity is frequent practice. But, in our opinion, it has certainly not been abandoned by either psychology or the business disciplines.

Common methods bias is reflected in MTMM when measurement items reflecting different latent constructs are correlated. There are no hard and fast guidelines regarding the extent to which these items may be correlated before concluding that common methods bias is a problem [Campbell and Fiske, 1959, Marsh and Hocevar, 1988].

A case from IS research will help to illustrate this threat. In studies of TAM, some researchers appear not to randomize questions dealing with the constructs of perceived usefulness, perceived ease-of-use, and system usage. As a result of this methodological artifact, respondents may be sensing the inherent constructs via the ordering of questionnaire items and they may respond accordingly.[17] In Table 2, each column represents a possible ordering of questionnaire items of

Table 2. Item Ordering Threats to Construct Validity through Common Methods Bias

Non-Randomized Presentation of Items		Randomized Presentation of Items	
1	I am very likely to try out CHART-MASTER.	2	I will very likely use CHART-MASTER.
2	I will very likely use CHART-MASTER.	4	Using CHART-MASTER in my job would enable me to accomplish tasks more quickly.
3	I will probably use CHART-MASTER.	5	I expect my company to use CHART-MASTER frequently.
4	I intend to use CHART-MASTER.	8	I would find it easy to get CHART-MASTER to do what I want it to do.
5	I expect my company to use CHART-MASTER frequently.	9	Using CHART-MASTER would make it easier to do my job.
6	Using CHART-MASTER in my job would enable me to accomplish tasks more quickly.	11	I would find CHART-MASTER easy to use.
7	Using CHART-MASTER would improve my job performance.	15	I am very likely to try out CHART-MASTER.
8	Using CHART-MASTER in my job would increase my productivity.	18	Using CHART-MASTER would improve my job performance.
9	Using CHART-MASTER would enhance my effectiveness on the job.	19	It would be easy for me to become skillful at using CHART-MASTER.
10	Using CHART-MASTER would make it easier to do my job.	21	Using CHART-MASTER in my job would increase my productivity.
11	I would find CHART-MASTER useful in my job.	23	Learning to operate CHART-MASTER would be easy for me.
12	Learning to operate CHART-MASTER would be easy for me.	31	My interaction with CHART-MASTER would be clear and understandable.
13	I would find it easy to get CHART-MASTER to do what I want it to do.	34	I would find CHART-MASTER useful in my job.
14	My interaction with CHART-MASTER would be clear and understandable.	35	I will probably use CHART-MASTER.
15	I would find CHART-MASTER to be flexible to interact with.	40	I would find CHART-MASTER to be flexible to interact with.
16	It would be easy for me to become skillful at using CHART-MASTER.	41	I intend to use CHART-MASTER.
17	I would find Chartmaster easier to use.	46	Using CHART-MASTER would enhance my effectiveness on the job.

the original Davis instrument [1989]. Column 1 presents the non-random ordering of the items that has characterized some TAM research. It is clear that even individuals unfamiliar with TAM and its basic hypotheses could easily infer that items 1 to 5 in column 1 are related (usage measures) as are items 6 to 11 (perceived usefulness measures) and items 12 to 17 (perceived

[17] Lack of random ordering of items may also explain some of the extremely high Cronbach alphas in the upper 0.90 range found throughout the TAM research stream. Cronbachs of greater than .95 are highly suspicious, for this reason.

ease-of-use measures). In short, the method itself is likely contributing to the pattern of responses rather than revealing the underlying, so-called "true" scores.

Conversely, column 2 shows a randomized presentation of items where, judging from the range of the numbering, we can see that other TAM-unrelated items must also be appearing in the instrument. Thus, the reason for randomized presentation is to minimize mono-method or common methods bias [Cook and Campbell, 1979], which is a threat to both discriminant and convergent validity (and, as discussed below also a threat to reliability).

Although randomizing items may reduce methods bias, a careful reading of Campbell and Fiske [1959] suggests that common methods bias can even be a problem when steps are taken to separate construct-related items randomly. It takes little stretching of the imagination to see how a respondent reading item 18 in Table 2 would naturally associate it with items 21 and 46 since the items, which utilize the same anchors, are still within the same instrument. Again, the method itself may be a major factor in how participants respond rather than a careful, thoughtful response that reveals the true score.[18]

A method of assessing common methods bias is a second order CFA in LISREL. This method can be used to assess common methods bias even when only one data collection method is used so long as data are collected at different points in time, such as when the same instrument is administered to the same population at different times. The second order CFA should be constrained so that there is one latent construct for each combination of method (or time when the questionnaire was administered) and trait (measures). These measures compose the first order factors. Second order factors are then created to represent each of the methods and each of the traits. The CFA is then constrained so that each first order factor loads on two second order factors representing its method and its trait, respectively. The correlations between the second order latent constructs representing methods and the second order latent constructs representing traits are set to zero, meaning that while methods and traits are allowed to correlate among themselves, they are not allowed to correlate with one another [Marsh and Hocevar, 1988]. This technique is superior to MTMM because it does not assume *a priori* that the measurement model that the researcher specified is necessarily the most valid one [Marsh and Hocevar, 1988]. Applying this technique, researchers can assess the significance of common methods bias by simply collecting the data at several points in time and running a second order CFA. Nevertheless, second order CFA is extremely rare in MIS research.

Woszczynski and Whitman [2004] found that only 12 of 428 articles in the IS literature over the period 1996-2000 even mentioned common methods bias. They list the means by which IS authors avoided this threat, particularly multiple methods of gathering independent and dependent variables, but the fact remains that few articles using one method test for the presence of this bias.

In the final analysis, the best heuristic for dealing with common methods bias is to avoid it completely to begin with. The use of maximally different methods for gathering data is a superior approach to testing for bias in data gathered with the same method [Campbell and Fiske, 1959]. It is especially desirable to apply a different method for dependent measures than independent measures [Cook and Campbell, 1979]. Perceptual data for independent variables could be counterpointed with archival data for dependent variables, for example. In this situation, there is

[18] Methodologists proposed quantitative techniques for analyzing single methods bias such as posed in this case. Avolio et al. [1991] applied WABA (within and between analysis) to determine whether methods variance exists when two or more constructs are measured through information from a single source (method). Bagozzi and Phillips [1982] and Bagozzi [1980] use SEM to determine the extent of common methods variance via a single source. Clearly, though, the most logical and safest way to determine if methods bias is a problem is to have more than one method in order to be able to compare the results *a la* the techniques proposed by Podsakoff et al. [2003].

no possibility that the use of the self-report method for the independent variables could influence the archival data gathered independently for the dependent variables.

Heuristics for Construct Validity

It is obvious from our lengthy discussion of construct validity, that this scientific check is one of the most critical procedures a researcher can perform. Without knowing that constructs are being properly measured, we can have no faith in the overall empirical analysis. Many established techniques are available for asserting valid constructs, and many more are evolving. What would seem to be best practice at the present time is to use one or more techniques for testing discriminant and convergent validity, including factorial validity. Because these approaches are reasonably well understood, we would argue that establishing construct validity should be a *mandatory* research practice, in general. In addition, as argued above, common methods bias can be avoided by gathering data for the independent variables and dependent variables from different sources, or, if a single method is used, to test it through SEM. Testing for common methods bias is a *highly recommended* technique, therefore. Nomological validity is likewise a *highly recommended* technique, to be thought of as supplemental to conventional construct validity approaches.

PREDICTIVE VALIDITY

Also known as "practical," "criterion-related," "postdiction," or "concurrent validity,"[19] predictive validity establishes the relationship between measures and constructs by demonstrating that a given set of measures posited for a particular construct correlate with or predict a given outcome variable. The constructs are usually gathered through different techniques. The purpose behind predictive validity is clearly pragmatic, although Bagozzi and Fornell [1982] argue that the conceptual meaning of a construct is partly attributable to its antecedents and consequences.

Figure 5 illustrates the key elements in predictive validity. Construct A or the independent variable, also known as the predictor variable, is thought to predict construct B or the dependent variable, also known as the criterion variable. The goal is simply prediction. It is not necessary to provide evidence of a theoretical connection between the variables. In the case where the theoretical connection is recognized, predictive validity serves to reinforce the theory base [Szajna, 1994].

The widespread use of GMAT scores to predict performance in graduate studies is a case in point in the academic setting. Decision-makers implicitly believe that constructs about mathematical or verbal ability will lead to higher performance in management graduate school and use a given instrument like the GMAT for highly practical reasons. Evidence that GMATs predict student performance well [Bottger and Yetton, 1982, Marks et al., 1981] suggests that the GMAT instrument demonstrates good predictive validity. As discussed in Nunnally [1978], predictive validity differs from a simple test of a model or theory in that it does not require theoretical underpinnings.

[19] Nunnally [1978] remarks that postdiction, concurrent, and predictive validity are essentially the same thing except that the variables are gathered at different points in time. Campbell [1960] discusses the practical nature of this form of validity. Rogers [1995] considers predictive and concurrent validities as subtypes of criterion-related validity.

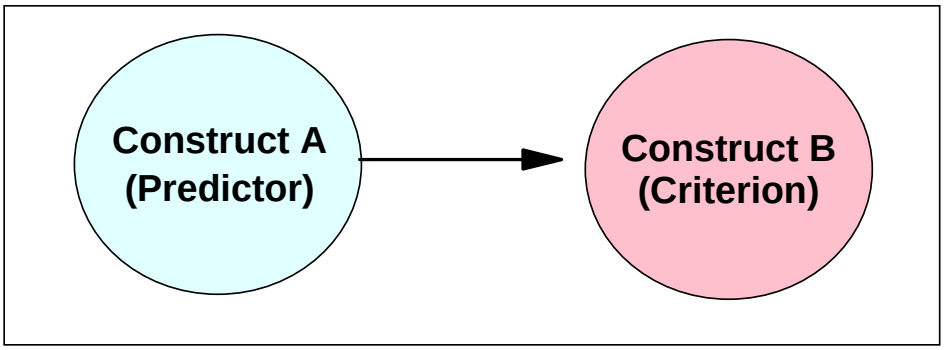


Figure 5. Pictorial Model of Predictive Validity

Campbell [1960], Cronbach [1990], and Bagozzi [1980] all conceptualize predictive validity as separate and distinct from content and construct validity. Other methodologists believe that it may be an aspect of construct validity in that successful predictions of links of constructs to variables outside a theoretical domain also validates, in a sense, the robustness of the constructs [Mumford and Stokes, 1992], as in the case of nomological validity. Yet, in that predictive validity does not necessarily rely on theory in order to generate its predictions, it is clear that it does not have the strong scientific underpinnings that arise from basing formulations of constructs and linkages on law-like principles. Although not discussed in Straub [1989], predictive validity could be put to better use in IS research,[20] especially in circumstances where it is desirable to show the applied value of our research [Cronbach and Meehl, 1955].

Vignette #6: Example of Predictive Validity

A good example of predictive validity can be found in Szajna [1994] prediction of choice of a system through criterion TAM constructs. The dependent variable, choice of system, served a pragmatic purpose. In this study, perceived usefulness and perceived ease-of-use were measured at one point in time. They were used to predict actual choice of a database management system to be used in an academic course at a later time. By varying the dependent variable from the traditional theoretical outcome in the TAM literature, i.e., intention to use/system usage, to system choice, Szajna [1994] was able to validate both the exogenous and endogenous constructs. In her analysis, TAM constructs proved to be accurate predictors 70% of the time, which, based on a z-score analysis, was highly significant over a chance prediction.

Heuristics for Predictive Validity

While the use of research constructs for prediction serves the practitioner community, it is generally not conceived of as being necessary for scientific authenticity. For this reason, we categorize it as an *optional* practice.

RELIABILITY

While construct validity is an issue of measurement *between* constructs, reliability is an issue of measurement *within* a construct. The concern is that instrument items selected for a given construct could be, taken together, error-prone operationalizations of that construct. Figure 6 shows that reliability of constructs A and C, being reflective constructs, is calculated based on the

[20] Predictive validity was considered to be part of construct validity in Boudreau et al. [2001].

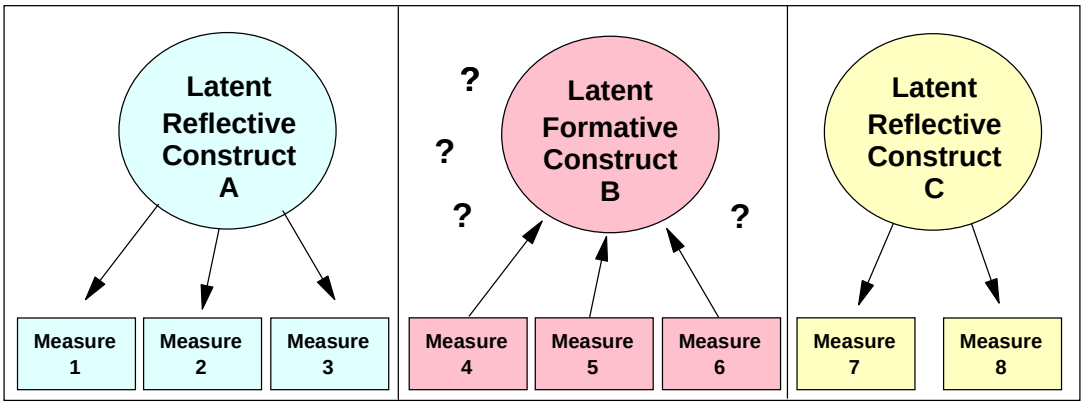


Figure 6. Pictorial Model of Reliability

extent to which the measures serparately correlate or move together. The reliability of one construct is independent of and calculated separately from that of other constructs, as depicted in the separate boxes in Figure 6.

Latent constructs that are formative may involve such different aspects of a construct that they do not correlate [Diamantopoulos and Winklhofer, 2001]. It is not clear, therefore, that reliability is a concept that applies well to formative constructs. These aspects "form" the construct, but do not necessarily "reflect" it in correlated measures.

As pointed out in Cronbach [1951], reliability is a statement about measurement accuracy, i.e., "the extent to which the respondent can answer the same questions or close approximations the same way each time" [Straub, 1989]. The philosophical underpinnings of reliability suggest that the researcher is attempting to find proximal measures of the "true scores" that perfectly describe the phenomenon. The mechanism for representing the underlying reality is integral to all data and data gathering [Coombs, 1976].

Six generally recognized techniques are used to assess reliability:

1. internal consistency,
2. split halves,
3. test-retest,
4. alternative or equivalent forms,
5. inter-rater reliability, and
6. unidimensional reliability.

where techniques 2 through 4 are considered to be "traditional" methods and 1, 5, and 6 are more recently employed techniques. Assessed through statistical packages such as SPSS and SAS, composite reliability coefficients analogous to internal consistency are also available in SEM. In addition, covariance-based SEM, such as LISREL, can be used to assess unidimensionality.

Boudreau et al., [2001] assessed the extent to which reliability was considered by IS researchers when developing their instruments. They discovered that the majority, that is 63%, assessed reliability. Most of this work (79%) estimated reliability through the standard coefficient of internal consistency, i.e.,Cronbach's α. Only in rare cases were other methods been used to verify the reliability of measures. Specifically, 2% used test/retest, 2% used split halves, and 21% used inter-coder tests. Moreover, the use of more than one reliability method occurred in only 13% of the studies assessing reliability. This state of affairs is regrettable because the use of additional methods to calculate reliability, or a combination of methods, would strengthen this component of instrument validation. The following subsections discuss the six types of reliability.

Internal Consistency

Internal consistency typically measures a construct through a variety of items within the same instrumentation. If the instrument in question is a questionnaire, items are varied in wording and positioning to elicit fresh participant responses. Moreover, if the scores from each of these items correspond highly with each other, the construct can be said to demonstrate acceptable reliability.

Reliability is the statistic most often used to evaluate internal consistency. This statistic is sensitive to the number of items forming the scale, so that a large number of items, say ten or above, will often yield high alphas, even if some measures are error-prone and not highly related to the other measures, i.e., reliable. Cronbach's α assumes that all items being considered for each construct are identically scored, as, for example, through Likert scales. If this assumption is not met, the researcher may plug Spearman correlations into the K20 formulation[21] or, more closely related to the likely values of a Cronbach's α, use reliability statistics generated by or calculated from SEM such as LISREL or PLS. Alternatively, some software packages such as SPSS 10.0 make such adjustments automatically.

What is ironic in assessing reliability is that high values (0.95 or greater) are more suspect than those in the middle alpha ranges.[22] When respondents are subjected to similar, identical, or reversed items on the same instrument, it is possible that very high reliability values simply reflect the ability of the participants to recall previous responses, which suggests that they were not responding naturally to the intent of each question to elicit their underlying true score. In short, the method itself became an artifact in the measurement. This threat of common methods bias is discussed at greater length in Campbell and Fiske [1959].

How can IS researchers fairly test their reliabilities and reduce the threat of common methods bias at the same time? Internal consistency testing is most valid when the instrument items are randomized or, at least, distributed in such a manner that the respondent cannot, in effect, guess the underlying hypotheses [Cook and Campbell, 1979]. Assume for the moment that all nine items measuring a single construct are arranged in sequential order on the instrument. In this scenario, it is extremely likely that the method itself (the side-by-side arrangement of the items) would determine the scores to a large extent and that a very high reliability statistic will result whether the item values truly represent the respondent's assessment or not. A robust arrangement of the instrument, therefore, is to separate items so that common methods bias is minimized.

Vignette #7: Example of Internal Consistency Reliability

Nearly all IS researchers prefer internal consistency statistics for reliability testing. In Grover et al.'s [1996] study of IS outsourcing, values for the major constructs ranged from .89 to .97. Using previously validated scales to measure media social presence, Straub [1994] used multiple items to examine the social presence of E-Mail and FAX as perceived by both American and Japanese knowledge workers. Cronbach's αs were .83 and .84. These values are acceptable, according to Nunally's rule of thumb, which allow values as low as .60 for exploratory research and .70 for confirmatory research [Nunnally, 1967].[23]

---

[21] The formula for coefficient α is: (k/k-1) (1- (Σ σ_i^2)/(σ_t^2), where k = number of parts/items in the scale, σ_i^2 = the variance of item i, and σ_t^2 = the total variance of the scale.

[22] Meehl [1967] makes a similar case for situations where the corroboration of social science propositions becomes weaker as precision gets better.

[23] Nunnally [1978, 1994] reassessed these guidelines in his subsequent articles, but his 1967 guidelines can be taken as reasonable if one allows more latitude for exploratory work.

Internal consistency testing is always subject to the charge of methods bias, but this technique at least reduces the impact of common methods bias on the true scores. For example, many researchers used Davis' TAM instrument without randomly varying the arrangement of the items composing each scale (Table 2, and discussion of discriminant validity). Non-randomized ordering of questionnaire items could easily explain the high Cronbach's αs that are reported as well as other methods artifacts related to discriminant and convergent validity [Straub et al., 1995].

Traditional Tests for Reliability

Straub [1989] did not discuss some traditional tests for reliability, including split-half, test-retest, and alternative forms [Cronbach, 1971, Parameswaran et al., 1979]. Although Parameswaran et al. [1979] criticize these traditional reliability tests for the assumptions in their theoretical underpinnings, IS researchers need to understand the basic approach in these techniques to be able to intelligently review manuscripts that choose to use these tests. In some specific cases these techniques continue to be useful in IS research, and IS researchers need to understand why these cases justify the use of these techniques.

Split Half Approaches. A traditional form of reliability assessment is split half testing. In this procedure the sample is divided into equal sub-samples and scores on the halves correlated. With these correlations a reliability coefficient can be obtained [Nunnally, 1978] by using the average correlation between items, as in all reliability estimating. Nunnally [1978] points out that the main difficulty with this technique is that different results are obtained depending on how one splits the sample. A random splitting will result in different correlations than an even-odd splitting, for example. Hence, the ability of the instrumentation to reflect true scores is not clearly and unambiguously estimated by the method. Moreover, if enough different splits are made, the results approximate Cronbach's coefficient alpha ([Nunnally, 1978] p. 233). A special purpose is needed for using this technique, as in the case of Segars [1997], discussed later. In general, its use is subsumed by internal consistency tests.

McLean et al. [1996] measured the importance of above-average salaries to IS graduates as they progress through the early months of their IS careers. As part of this study, they test the reliability of their Job Satisfaction scale through split-half reliability. Scores ranged from .47 to .89, with an overall mean of .80. Another part of their instrument was an Organizational Climate scale, which was tested via Cronbach αs ranging from .62 to .90.

Test-Retest. Test-retest approaches to determining whether an instrument will produce the same scores from the subjects every time is a form of reliability testing that can be used effectively in certain circumstances [Cronbach, 1951, Nunnally, 1978, Nunnally and Bernstein, 1994, Peter, 1979]. Test-retest involves administration of the instrument to the same sample group twice, the second administration being typically after a one or two week interval [Peter, 1979]. One assumption underlying this test is that if the instrument is reliable, the intervening time period will not result in widely different scores from the same subject and measurement error will be low.

A good example of the use of test-retest is Hendrickson et al. [1993]. In this test of the reliability of TAM measures over time, the instrument was administered to 51 subjects using a spreadsheet package and 72 subjects using a database management package. The same test was administered to the subjects after a three day period. Reliability values were comparable, albeit slightly lower than Davis' [1989]. What is clear in the case of this validation research is that reliability of the TAM instrument was well established before, but via the administration of a single instrument (i.e., internal consistency estimates). Examining the stability of the scales over time, therefore, was a valuable validation exercise.

Clearly, there are several inevitable threats in the use of this technique.

- The test-retest threat [Cook and Campbell, 1979]. That is to say, the answers may be similar because a respondent simply recalls the previous answer and not because the second score necessarily verifies the accuracy of the first score. In all likelihood,

this threat is no more or less problematic than the methods bias threat for internal consistency. In general, the longer the time between administrations of the instrument, the less likely it is that the participant will remember the prior responses [Rogers, 1995], and, therefore, the lower is the test-retest threat [Hendrickson et al., 1994].

- Lengthening the time interval, however, raises another threat, that of an intervening event legitimately affecting the true score [Peter, 1979]. In such a case, it is not possible to distinguish between reliability and causality. Peter [1979] discusses other threats that IS researchers need to be aware of.

Alternative or Equivalent Forms As discussed in Peter [1979] and Nunnally [1978], alternative forms involve comparisons between the scores for various constructs as represented by the instrument and scores in other "tests" or instruments. For example, a sample group tested for computer literacy scores using one instrument can be compared to similar scores on a related instrument. Alternative forms have the same problems as test-retest in that they are administered at different points in time. Moreover, the reliabilities computed for different alternative tests could vary significantly. Which of these tests represents the better comparative test cannot be determined *a posteriori.* Alternative forms have not been used recently in IS research. Boudreau et al.'s [2001] sampling of journal articles within a recent three year period did not reveal a single example of this form of reliability testing. The problem equivalent forms creates is obvious. There is little to go on with respect to best practice.

The other difficulty with this approach is that its procedures are not completely distinguishable from discriminant validity tests, which, again, assume that measures show high construct validity when they are able to differentiate between sets of variables, some of which are measuring highly correlated concepts [Cronbach and Meehl, 1955]. Given these difficulties, this form of reliability should not be a first choice for IS researchers.

Inter-Rater or Inter-Coder Reliability

Often, in empirical research, collected data does not manifest itself in a natural quantitative form. A great deal of unstructured and semi-structured discourse in interview transcripts data falls into this category. Even structured interview data, such as verbal responses to a scale provided to the interviewee, can be complicated by qualifications made by the respondent. In such cases, researchers find it desirable to code the data so that they can analyze it and interpret its underlying meaning.

Inter-rater reliability, in which several raters or judges code the same data, is of great interest, therefore, in both quantitative and qualitative research [Miles and Huberman, 1994]. Yet, in the context of this dual facility, several issues are unresolved. One question is whether the terms reliability and validity even apply to qualitative work [Armstrong et al., 1997, Denzin and Lincoln, 1994] or, if they are applicable, in which circumstances [Burrell and Morgan, 1979, Lacity and Janson, 1994]. The other major questions are whether the techniques produce accurate and reproducible results [Armstrong et al., 1997] or are suitable for all forms of data [Jones et al., 1983, Perreault and Leigh, 1989]. Miles and Huberman [1994] suggest coders need to be trained with definitions of key constructs and a process for developing consistent coding. Once properly coded, the data are then analyzed via several techniques.

Cohen's coefficient Kappa is the most commonly used measure of inter-rater reliability. Pearson's or Spearman correlations (including average correlation, interclass correlation, and the Spearman-Brown formula) as well as percentage agreement are sometimes used for the case of two raters [Jones et al., 1983].[24] Miles and Huberman's [1994], Landis and Koch's [1977], and Bowers and Courtright's recommendations [1984] for minimum inter-rater reliability are .70.

[24] Data comparing two raters can be reorganized by systematically transferring the higher of the ratings to the same field or column. Lacking this transformation, the reliabilities will be highly attenuated and

Vignette #8: Examples of Inter-Rater Reliability

In Lim et al.'s [1997] study of computer system learning, two independent coders scored tests based on explicit instructions. These instructions described how to determine if a good explanation was provided by the respondent, and were thus fairly detailed. A correlation of .84 was assessed. Before further statistical analysis was performed on the test scores, disagreements between the coders were reconciled.

Another example is Pinsonneault and Heppel [1997/98], who created an instrument to measure anonymity in groupware research. To create their scales, graduate students sorted items into categories believed to be the constructs of interest. Level of agreement among raters was established via percentages and coefficient kappa. Initially, the agreement score was 79% and the Kappa was .75, indicating good inter-respondent agreement.

Using the more stringent measurement of Kappa designed by Umesh et al. [1989], Boudreau et al. [2001] classified articles according to their use or non-use of research validities. They determined that the Kappa for their raters' coding exceeded the benchmark .70 threshold. Percentages of agreement were in the 74% to 100% range.

Unidimensional Reliability

Unidimensionality is an important statistical test, but, alas, is perhaps the least understood, newest, and certainly the least applied. Unidimensionality is a property of a measurement item that states and examines that the item measures, that it reflects, only one latent construct. Unidimensionality is assumed *a priori* in many measurements of reliability, including Cronbach's α. Gefen [2003] extensively discusses this relationship.

Techniques in covariance-based SEM can also help to determine the unidimensionality and the traditional reliability of a construct. Unidimensionality means that each measurement item reflects one and only one latent variable (construct) [Anderson et al., 1987, Gefen et al., 2000, Segars, 1997]. That is, it means that tests should not reveal that a measurement item significantly reflects more than the latent construct to which it is assigned. The terms frequently used to discuss this validity are: "first order factors," "second order factors," etc. A first order factor is the most macro level conceptualization of a construct. It is composed of more than one second order factors, which, together, would be reflective or formative of the first order construct [Gefen et al., 2000].

Unidimensional reliability is a relatively new, highly sophisticated approach for validating reliability, although it was long recognized as a basic assumption upon which other measures of reliability rely.[25] Unfortunately, prior to the advent of SEM and Item Response Theory [Hambleton et al., 1984], the measurement of unidimensionality was extremely laborious from a statistical standpoint. The rule for determining unidimensionality is that such constructs will not show "parallel correlational pattern[s]" ([Segars, 1997], p. 109) among measures within a set of measures (presumed to be making up the same construct) and among measures outside that set

---

inaccurate. Consider, for example, a case where the raters always differed by only one value on a five point scale, about half of the time in one direction and about the other half of the time in the other direction. The correlation between these raters using an untransformed dataset would be close to .00. If the data is reorganized in the manner suggested, the correlation is 1.0, reflecting the fact that the raters were tracking each other closely (consistently only a one point calibration difference). Calibration clearly remains an issue in the illustrative inter-rater dataset, but the reliability is certainly not as negligible as a .00 would indicate.

[25] It is unclear at this point as to whether unidimensionality is only a characteristic of reliability or whether it can also be or should be thought of as a form of construct validity. It is likely that it can be used in either or both contexts. What is probably more important is that IS researchers begin to work with this validation tool more frequently to gain a clearer picture of the internal structure of their measures and constructs.

(see also [Anderson et al., 1987]). As discussed in Long [1983a, 1983b] and Jöreskog and Sörbom [1993, 1994], and Chin [1998b], fundamental capabilities of SEM allow researchers to test the relationships between instrument items (measures, indicators, or observed variables) and latent variables (constructs). Researchers examine first order and second order models to determine if the posited structure of variables is unidimensional.

Vignette #9: Examples of Unidimensional Reliability

Segar's study [1997] of IT diffusion and IT infusion variables is an exemplar for how unidimensional reliability can be assessed. After CFA analysis investigated the two factor- ten item model for unidimensionality and measurement fit, two of the ten items were dropped and unidimensionality established. The resulting two factor model with 8 items was used to derive reliability statistics in two ways. A split-half technique generated an alpha coefficient for both the IT diffusion and IT infusion scale items. These values were acceptable at .91 and .87. Moreover, the average variance extracted by the items was .73 and .63, respectively, which was sufficiently above the .50 cutoff value mentioned above.[26]

Sethi and King [1994] used CFA to determine that there were seven unidimensional constructs in their "Competitive Advantage Provided by an Information Technology Application" (CAPITA) research instrument. Dropping two constructs that did not qualify, each of the seven factors was shown to be unidimensional.

Readers are urged to examine the tutorial by Gefen [2003]. This paper presents an example and step-by-step walkthrough on the use of unidimensionality in LISREL and the threats it addresses. It includes real data that can be used to replicate the arguments on the necessity of performing unidimensionality analysis. The tutorial includes a running example that shows how ignoring threats to unidimensionality can seriously affect conclusions drawn from the structural model. The tutorial also shows that these threats cannot be assessed with a PCA. Gefen [2003] used CFA to show the unidimensional nature of the Perceived Usefulness and Perceived Ease of Use constructs of TAM together with the SPIR (social presence/information richness) measures used by Gefen and Straub [1997].

Mono-Operation Bias

Cook [1979] point out the threat to reliability (a threat which also holds for construct validity) that results from mono-operationationalization of constructs. With such single item measures, we cannot be sure that we have captured the construct because we have no means to validate the metric. It may be right or wrong or somewhere in between, but we have no standard against which to judge the researcher's choice of item.

Nevertheless, there are cases where the researcher has little choice but to use a single measure for certain constructs. This may occur because of the intractability of the construct, the danger that respondents will become indifferent due to lengthy instrumentation, or a host of other legitimate reasons. In such cases, the researchers need to make a convincing argument for their measure, and balance this off with value in other parts of the study. In the long run, if there is a compelling logic to the theory and the posited constructs, measurement will be approached by other researchers. While this position may seem at odds with validation as a primary and prior

[26] The resulting two factor model with 8 items shows high fit values and also passes tests for both convergent and discriminant validity. Thus, the concept of unidimensionality clearly applies to both reliability and to construct validity.

process in the research process, we realize that the guidelines offered here are a Weberian "ideal type" which are rarely seen in their entireity in practice. They are not unobtainable, as we will later demonstrate, but most of our work falls short of attaining perfection or near perfection in validation.

Heuristics for Reliability

Reliability assures us that measures that should be related to each other within the same construct are, indeed, related to each other. Without reliable measures, it is difficult to see how the data can be trusted, any more than instruments lacking construct validity can be trusted. One form or another of reliability is *mandatory* for scientific veracity.

If qualitative data is not being coded, then internal consistency measures (Cronbach's αs or SEM internal consistency/composite statistics) should be used first in the development of instruments. Since other forms of reliability contain limitations, they may be applied in more mature research streams.

Unidimensional validity is an important new approach in the IS researcher's toolkit. Because its previous use by the field is limited, we suggest that it be classified as *optional* until we gain more experience with it and understand its capabilities better.

MANIPULATION VALIDITY

Manipulation validity (a.k.a. manipulation checks) is traditionally inserted into experimental procedures/tests to measure the extent to which treatments (IVs) are perceived by the subjects [Bagozzi, 1977]. As shown in Figure 7, manipulated constructs A and B, or treatments, are the

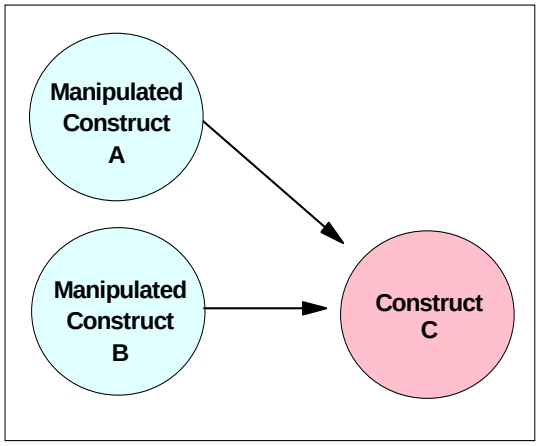


Figure 7. Pictorial Model of Manipulation Checks

independent variables hypothesized to produce an outcome. The manipulation validity is an assurance on the part of the researcher that the manipulations "took" in the subjects. In the case of physiological treatments, like new drugs, little doubt exists,[27] but considerable doubt is encountered in cases where researchers are manipulating the subjects' perceptions, through experimental tasks or exercises.

It needs to be understood that subjects must be aware of certain aspects of their manipulation, but not others. Clearly, subjects being asked to respond to a scenario manipulating high level of sunk costs in IT investments, as in [Keil et al., 1995], must be cognizant of this level for the

[27] Nevertheless, physiological researchers usually guard against the opposite effect. Subjects who did not receive a treatment (control group) are supplemented by a group given a placebo so the control group can be compared to those who do not "perceive" that they had no drug and, and therefore, no effect.

manipulation to create any impact. But it would be highly undesirable if these same subjects were able to guess [Argyris, 1979] the underlying "project escalation" hypotheses of this experiment and respond accordingly.[28] Manipulation validity is designed to ensure that subjects are, indeed, manipulated as intended. Therefore, it is a validity that can be examined empirically.

Manipulation validity can be simple and straight-forward or complex. One common form of check is a simple question or questionnaire item on the experimental test that asks the subjects directly if they experienced the manipulation. Or, it can be assessed in a more sophisticated way, using ANOVA, discriminant analysis, or other techniques [Perdue and Summers, 1986].

Manipulation validity is not assessed frequently enough in IS experimental settings. Indeed, Boudreau et al. [2001] report that only 22% of the field and laboratory experiments in their sample assessed this type of validity. As to the particular means by which manipulation validity was assessed, their sample showed that techniques such as t-test, χ^2, and ANOVA were deployed about twice as often as descriptive statistics such as counts, means, and percentages.

Vignette #10: Examples of Manipulation Validity

Keil et al. [1995] conducted manipulation validity in a straightforward way in their research. Their subjects responded true or false to whether or not they were given a choice of an alternative course of action, which was one of the treatments. In Simon et al. [1996], two manipulation checks were employed to assess perceptions of and reactions to the training treatments. The results of an ANOVA were that the three training treatments — (1) instruction, (2) exploration, and (3) behavior modeling — showed significant differences between groups on perceptions of training, but not on reactions to training.

Using a relatively more sophisticated technique, Gefen and Straub [1997] employed discriminant analysis to assess manipulation validity. Subjects were randomly assigned to one of five experimental groups. The baseline group examined a common Web-site; the four treatment groups examined various additions to this baseline Web-site. Students then answered twelve true/false manipulation check questions that tested their responsiveness to these variations of the Web-site. The success of the manipulation was assessed using a Multiple Discriminant Analysis (MDA) to examine whether the students could be reclassified into their original treatment groups based on the manipulation check questions. The MDA showed four significant canonical discriminant functions, as would be expected of five treatments groups, and correctly classified over two-thirds of the students. This percent of successful manipulation exceeds the 50% rule of thumb guideline suggested by Jarvenpaa et al. [1985].

Heuristics for Manipulation Validity

Manipulation validity is *mandatory* for nearly all types of experimentation. Without these checks, the experimenters cannot be certain which subjects were exposed to the treatments and which were not. When a subject is being treated with a physical substance, such as a drug, manipulation validity is not needed. But in the social science research that characterizes a great deal of management research, subjects may not be paying attention or may be uninterested in the experimental treatment. The manipulation validity is one way of attempting to purify the data collected by discriminating between those who truly received the treatment and those who did not. Practice varies, but removing unmanipulated subjects from the pool of data will generally improve significance of effects, and, for this reason, this heuristic is highly recommended. It

[28] Hypothesis-guessing is an experimental confound. Careful experimental procedures insulating the subjects from the hypotheses are designed to protect against the problem [Orne, 1962, Orne, 1969].

should be noted, however, that because unmanipulated subject responses presumably add unexplained variance, inclusion of these subjects in the dataset is a more robust testing of the hypotheses and some researchers may choose to retain them for this reason.[29] The danger of allowing unmanipulated subject responses to remain in the dataset is Type II errors, that is, concluding that this is no effect when, in fact, there was one. Removing unmanipulated responses helps to avoid Type II errors.

STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity assesses the mathematical relationships between variables, and makes inferences about whether this statistical formulation correctly expresses the true covariation [Cook and Campbell, 1979]. It deals with the quality of the statistical evidence of covariation, such as sources of error, the use of appropriate statistical tools, and bias. Type I and Type II errors are classic violations of statistical conclusion validity. IS field has also been able to take advantage of new techniques developed in the last decade that assist in establishing statistical conclusion validity. These techniques take different approaches to establishing whether, statistically, there is a "critical realism" ([Cook and Campbell, 1979], p. 29) in the relationship between variables or sets of variables. These tools are known as structural equation modeling (SEM) techniques and they are sufficiently different from bivariate, nonparametric, and multivariate techniques to call for special treatment in this paper.

The two types of SEM are covariance-based and PLS. Covariance-based SEM examine the entire matrix of covariances (or correlations, depending on how the model is run) including covariances that are not specified in the model. PLS, on the other hand, examines the proposed model alone, ignoring other covariance that is not explicitly stated in the model. Gefen et al. [2000] present a detailed discussion and comparison. The effective use of SEM in IS research is the underlying issue. These SEM techniques are now widely used in the top IS journals [Gefen et al., 2000].

Heuristics of Statistical Conclusion Validity

Vignette #11: Examples of Statistical Conclusion Validity

IS authors use statistical conclusion validity, specifically justifying the type of tool they use based on its inherent distribution assumptions and its ability to deal with small sample sizes. Indeed, it is difficult to see how much positivist, quantitative research could pass muster without it. In the next two examples the researchers, applying SEM, explain why they chose one statistical tool over others based on its statistical properties and distribution assumptions. Sambamurthy and Chin [1994] chose PLS, explaining at length why it is more appropriate than LISREL for their specific data and their predictive rather than confirmatory approach. Taylor and Todd [1995] preferred LISREL for their analysis because of its ability to compare alternative models dealing with well established theories.

Statistical conclusion validity is mentioned briefly here for the sake of completeness. This technique receives the single most attention in management research [Scandura and Williams,

[29] The counter argument to pooling the manipulated and unmanipulated subject responses is, of course, that the error terms of the unmanipulated subjects are unknown. Their responses to items cannot be assumed to be a random process and, therefore, they may just represent bad data that should be discarded. Contrariwise, eliminating subject responses lowers sample sizes, sometimes below the minimum of 20 per cell, commonly used as a heuristic, itself conservatively derived from minimums of 15 per cell for a student's T test.

2000]. The simplest way to document the heuristics in this category is to refer the reader to Gefen et al. [2000], where tables contain a complete set of heuristics.

III. GUIDELINES FOR RESEARCH PRACTICE

Although it improved over the years, instrument validation still needs to make major steps forward for scientific rigor in the field. Boudreau et al. [2001] call for "further heuristics and guidelines for bringing even more rigor to the process of positivist, quantitative research" (p. 13). Based on such observations and interpretations of prior work, validity rules of thumb can be expressed. These rules are essentially pragmatic measures indicating patterns of behavior that appear to be acceptable within the IS scientific community. No recognized means verifies the truth of such heuristics, other than through tradition, philosophical disputation, and evaluation of best of breed practice. It is traditional, for example, for IS researchers to use at least a .10 alpha level (Type I error) in their studies. Even in this case, it is more often than not the practice that .10 is associated with exploratory work whereas confirmatory work uses either a .05 or .01 alpha protection level. The numbers mentioned here represent what the community is willing to accept as a level of risk in statistical conclusion validity. If the IS community were suddenly willing to accept a 25% chance (.25) that the results being reported could be false, then a new alpha level would become the rule of thumb. The choice of levels cannot be established by mathematical or other means [Nunnally, 1978, Nunnally and Bernstein, 1994].

The same logic applies to statistical power, correlation values and explained variance, and a host of other statistical concepts. On the issue of statistical power, for instance, Cohen [1977] makes a case that .80 is reasonable for a medium effect size, given the history of values reported in the literature. Thus, this community standard implies that researchers should be willing to accept a 20% chance of false positives for medium effect sizes, and less so for large effect sizes.

With respect to a rule of thumb for correlation coefficients, Cohen [1988] argues that since the overwhelming majority of social science studies report relationships that correlate significantly at .50 or below, then a large effect is approximately .50, a moderate effect is .30, and a small effect is .10. Large effects, moreover, are likely so obvious as to be trivial whereas small effects are merely significant from a statistical, rather than practical point of view. Again, such heuristics are argumentative and could be challenged at any point by the scientific community.

The range of correlations in the published TAM stream is from 20-60%, and the acceptability of the explanatory power of any of these models is solely dependent on the judgment of the reviewers. Falk and Miller [1992] argue that a minimum of 10% explained variance is acceptable for scientific advancement.

Rules of thumb are desirable because of their practicality. Researchers can use them as *de facto* standards of minimal practice and as a first approximation for how true their instrumentation is, in reality. Any heuristic can be challenged by members of the community, and, with effective persuasion, a new rule of thumb then set. The summary list of our heuristics, all subject to challenge by the IS community, is presented in Tables 3 through 6.

Table 3. Mandatory Validities

Validity Component	Technique	Heuristic	Source
Construct Validity ❑ *Discriminant validity*	MTMM	Relatively low number of matrix violations; SEM estimates of error attributable to method.	[Campbell and Fiske, 1959] [Bagozzi, 1980]
	PCA	Latent Root Criterion (eigenvalue) of or above 1, although using a Scree Tail Test criterion is also accepted, in which case factors are accepted until the eigenvalue plot shows that the unique variance is no longer greater than the common variance. Loadings of at least .40 (although some references suggest a higher cutoff); no cross-loading of items above .40. Items that do not load properly may be dropped from the instrument [Churchill, 1979].	[Hair et al., 1998]
	CFA as used in SEM	GFI > .90, NFI > .90, AGFI > .80 (or AGFI > .90, in some citations) and insignificant χ^2, combined with significant t-values for item loadings.	[Hair et al., 1998] [Segars, 1997] [Gefen et al., 2000]
❑ *Convergent validity*	MTMM	Significant homomethod, homotrait correlations.	
	PCA	Eigenvalues of 1; loadings of at least .40; items load on posited constructs; items that do not load properly are dropped.	[Hair et al., 1998]
	CFA as used in SEM	GFI > .90, NFI > .90, AGFI > .80 (or AGFI > .90, in some citations) and preferably an insignificant χ^2; item loading should be above .707 so that over half of the variance is captured by the latent construct; also, the residuals (item variance that is not accounted for by the measurement model) should be below 2.56.	[Hair et al., 1998] [Thompson et al., 1995] [Chin, 1998b] [Segars, 1997] [Gefen et al., 2000]
❑ *Factorial validity*	PCA	See PCA above for discriminant and convergent validity.	
	CFA as used in SEM	See CFA & SEM above for discriminant and convergent validity.	

Table 4. Mandatory Validities (Where Appropriate)

Validity Component	Technique	Heuristic	Source
Reliability ❑ *Internal consistency*	Cronbach's α; correlations; SEM reliability coefficients	Cronbach's α should be above .60 for exploratory, .70 for confirmatory; in PLS, should be above .70; in LISREL, EQS, and AMOS, should also be above .70.	[Nunnally, 1967] [Nunnally, 1978] [Nunnally and Bernstein, 1994] [Peter, 1979] [Thompson et al., 1995] [Hair et al., 1998] [Gefen et al., 2000]
❑ *Inter-rater reliability*	Coefficient kappa; correlations; percentages;	Coefficient Kappa > .70.	[Landis and Koch, 1977] [Miles and Huberman, 1994]
Manipulation Validity	Percentages; T-tests; discriminant analysis; ANOVA	Although no clear thresholds exist, higher percentages are clearly better; tests of significance; subjects who are not successfully manipulated should (arguably) be withdrawn from the dataset.	[Perdue and Summers, 1986]

Table 5. Highly Recommended Validities

Validity Component	Technique	Heuristic	Source
Content Validity	Expert panels or judges	High degree of consensus; judgmental except for content validity ratios computed using Lawsche.	[Lawshe, 1975]
Nomological validity	Comparison with previous nomological networks; regression; correlations; SEM	Comparisons with previous magnitude measures, e.g., path coefficients; also with previous variance explained.	
Common methods bias / Method Halo	Collect data at more than one period; collect data using more than one method; separate data collection of IVs from DVs	Run second order CFA to check for method bias.	[Marsh and Hocevar, 1988] [Cook and Campbell, 1979]

Table 6. Optional, but Recommended Validities

Validity Component	Technique	Heuristic	Source
Predictive validity	Z-scores; correlations; discriminant analysis; regression; SEM	Explained variances in the .40 range or above are desirable.	
Reliability ❑ *Split half*	Same as internal consistency	Cronbach's α>.60/.70 and < .95.	
❑ *Test-retest*	Same as internal consistency	Cronbach's α>.60/.70 and < .95.	
❑ *Alternative forms*	Same as internal consistency	Cronbach's α>.60/.70 and < .95.	
❑ *Unidimensional reliability*	Model comparisons	Model comparisons favor unidimensionality.	[Segars, 1997] [Gefen et al., 2000] [Gefen, 2003]

What our slow progress toward rigorously validated instruments suggests is that the guidelines for IS research practice may need to be strengthened as well as broadened to include validities discussed in this article. The 1989 Straub guidelines will, therefore, be subsumed into "Guidelines for the Year 2004 and Beyond," immediately below.

GUIDELINES FOR 2004 AND BEYOND

Two broadly stated guidelines emerge from the present study: research validities and innovation in instrumentation. Table 7 lists the research validities and indicates the recommendation for them.

Table 7. Guidelines for Research Validities

Validity	Recommendation
Content validity	Highly recommended
Construct validity	Mandatory
Predictive validity	Optional
Reliability (internal consistency)	Mandatory (where appropriate)
Reliability (split halves)	Optional in mature research streams
Reliability (alternative forms)	Optional in mature research streams
Inter-rater reliability	Mandatory (where appropriate)
Unidimensional Reliability	Optional
Manipulation validity for experiments	Mandatory (where appropriate)
Nomological validity	Highly recommended
Common methods bias	Highly recommended
Statistical conclusion validity	Mandatory

Instrumentation	Recommendation
Use of previously validated instruments	Highly recommended
Creation of newly validated instruments	Highly recommended

Research Validities

Gatekeepers at journals — both editors and reviewers — should require separate article sub-headed sections for validation of instruments, data-gathering approaches, and/or manipulations, as relevant; and, at the very least, insist on the standards in Table 7 for validation. Our logic, in brief, for the most common of these validities, follows.

Content validity

Highly recommended. Establishing content validity is a highly desirable practice, especially in the absence of strong theory and prior empirical practice specifying the range and nature of the measures.

Construct validity

Mandatory: Establishing construct validity (convergent and discriminant validity) is a necessary practice, with factorial validity being minimally required. For mature research streams, convergent and discriminant validity established through MTMM is recommended, as is nomological validity and ruling out common methods bias.

Predictive validity

Optional: Establishing predictive validity is useful for mature research streams.

Reliability

Mandatory: Establishing reliability is a necessary practice, with Cronbach's α tests being recommended over other tests of reliability. When LISREL or PLS are used, reliabilities generated by or calculated from these SEM techniques should replace (or at least augment) the Cronbach's scores. Internal consistency reliability should be the first generation test of an instrument; other types of reliability testing, such as test-retest should follow as the research stream matures. Where appropriate, inter-rater reliability is mandatory.

Unidimensional reliability

Optional in spite of its growing importance: At present, the technique is little known in IS research. Over time it could become mandatory because all reliability measures, including Cronbach's α, assume *a priori* that the measures are unidimensional. As IS researchers gain more experience with unidimensional reliability testing, this form will likely earn greater prominence.

Manipulation validity for experiments

Mandatory: Establishing manipulation validity is a necessary practice for determining the validity of treatments (independent variables) in experimentation. For other aspects of instrumentation, experiments should be subject to the same validity standards as other research methods.

Statistical conclusion validity

Mandatory: Establishing statistical conclusion validity is essential for all quantitative, positivist research.

Innovation in Instrumentation

Use of previously validated instruments

Highly recommended**:** For the sake of efficiency, researchers should use previously validated instruments wherever possible, being careful not to avoid previous validation controversies or to make significant alterations in validated instruments without revalidating instrument content, constructs, and reliability.

Creation of newly validated instruments

Highly recommended: Researchers who are able to engage in the extra effort to create and validate instrumentation for established theoretical constructs (nomological validity) are testing the robustness of the constructs and theoretical links to method/measurement change (see Boudreau et al. [2001], for more detailed argumentation). This practice, thus, represents a major contribution to scientific practice in the field.

Laboratory and Field Experiments and Case Studies

Laboratory and field experiments, as well as case studies, lag behind field studies with respect to most validation criteria [Boudreau et al., 2001]. This result is disheartening in that laboratory experiments are superb ways to test existing theory and new theoretical linkages. The field needs the rigor of internal validity that lab experiments bring to the overall mix of our research. As to positivist case studies, they are interesting because they are more likely to provide better qualitative evidence that instruments are scientifically valid [Campbell, 1975]. One would hope that an analysis of the state of the art in IS validation in the next decade would reveal large scale improvements, no only among field studies but also among laboratory experiments, field experiments, and positivist case studies.

Contingent Applicability of the Guidelines by Research Design

Some would argue (and we would be receptive to this point of view) that there are situations in which the validities should be applied differentially. Not all research designs are equal, after all.

Where would these occur? We invite the IS community to add to the table we offer below (Table 8), but it needs to be kept in mind that this table is only an initial attempt to spell out some of the conditions where the guidelines may need to be adapted to particular research situations. What is clear in the line of reasoning we pursued in this article is that the kind of highly theoretical, confirmatory work that most frequently appears in top ranked IS journals will require the rigor of full validation. The heuristics we proffer here are those that will lead to the requisite rigor that these journals should welcome.

Many IS journals are also open to exploratory work, and it is possible to argue cogently that a research design that was probing into new territory, where the theories of the field or contributing/reference fields may not apply, calls for a different set of validities. In Table 8, we suggest that content validity may still be applicable since the researchers are exploring new definitions of constructs and are implicitly ruling out some possible measures and ruling in others. Their definitional boundaries are of great interest. Empirical tests are of even greater value.

By the same token, exploratory work may not require the more exacting and comprehensive tests of construct validity. Readers might expect to see factorial validity and Cronbach's internal consistency reliability tests in this case.

Exploratory work may not yet be testing the strength of relationships between constructs. Therefore, predictive validity or nomological validity are beyond what many readers would require. Another possible set of contingencies could be added for matching or fitting the research design (including choice of validities to stress) to the research question. A seminal work that takes this point of view is Jenkins [1985]. It may be the case that certain research questions call for methodological approaches that are not covered in our philosophy. We fully recognize that this lack could be an oversight in the current paper. We once again encourage IS researchers to think along these lines and present alternative points of view.

Table 8. Contingencies for Where Validities Apply

No.	Positivist Design Contingency	Description	Validities Stressed
1	Exploratory work	Probing new areas that are not now well understood; these areas may not have strong theory bases from contributing or reference disciplines.	Content validity; straight-forward and initial factorial tests of construct validity and internal consistency reliability tests
2	Intractable Domains	Working in areas of study that are intractable or extremely difficult to measure; the standard should be that the field is better off with these insights and weak measures than without the results of the study.	Content validity; mono operations, but with explanations for the reasonableness of the measures
3	Confirmatory research in well established research streams	Confirming the relationships between constructs that have been found again and again in highly related streams.	All validities are likely applicable, especially nomological validity
4	Theoretical work	Simply testing theory or proposing refinements to theory and then testing them; relationships between constructs are the central elements in this kind of work.	All validities are likely applicable, especially nomological validity
5	Non-theoretical work	Examining the nature of a phenomenon through descriptive statistics primarily, or with highly exploratory hypothesis testing.	Content validity;[30] predictive validity
6	Previously validated instrumentation	Applying the validated instrumentation to a new phenomenon or, less commonly, in a replication.	All validities are likely applicable, but in much less detail than would be called for if the instrumentation were new; variant forms of reliability other than internal consistency would be appropriate
6	New instrumentation	Inventing new measures and procedures in cases where the theory was not well advanced or where it was advanced, but the prior instrumentation is weak	All validities are likely applicable, and in great detail; this is the heart of the demonstration of the usefulness of the new instrumentation

IV. MODEL OF INSTRUMENT VALIDATION

To demonstrate that these guidelines are by no means impossible or out-of-reach for many or even most IS researchers, a single example of how new instruments can be developed is offered

[30] It should be noted that descriptive work can be valuable to researchers seeking to validate the measures they are using.

next. Smith et al. [1996] validated their information privacy instrument through a judicious choice of most of the validation techniques we discussed. Each of these techniques are described briefly, to show the extent of the validation undertaken.

CONTENT VALIDITY

First, four different groups were asked to assess the content of dimensions that the authors proposed for information privacy.

Group 1: Three experts in the privacy area were given 72 preliminary questionnaire items.

Group 2: A reduced set of 39 items were then evaluated by 15 faculty and doctoral students. In order to compare responses, this group was split, with roughly one half receiving definitions of the sub-construct dimensions and the other, not.

Group 3: Thirty-two remaining items were next evaluated by 15 corporate employees.

Group 4: A focus group of 25 persons. The final scale included 20 items.

In that information privacy affects many groups, the use of judges from these different groups verified that the "content" of the items was likely not idiosyncratic or biased.

PRETEST

This resulting 20-item scale was further refined through administration to 704 bank, insurance, and credit card issuer employees. Exploratory factor analysis and reliability tests found support for most of the posited sub-constructs. Additional exploratory factor analysis with three revised versions of the instrument sampling information systems managers and graduate business students resulted in a 15-item instrument. Four subscales remained: Collection (4 items), Errors (4 items), Unauthorized Secondary Use (4 items), and Improper Access (3 items).

UNIDIMENSIONAL RELIABILITY

Using samples described immediately above, the authors next attempted to determine whether the hypothesized model of four dimensions offered the best fit to the data. Using the CFA capabilities of LISREL, four theoretically plausible alternative models (a unidimensional model, a three-dimensional model, a model with two main factors and three sub-factors, and the hypothesized four factor model) were compared. LISREL statistics indicated that the four sub-construct model was the best fit to the data.

CONSTRUCT VALIDITY

Both convergent and discriminant validity were assessed with a new sample of 147 graduate business students. LISREL statistics were used to determine that the sub-constructs converged (viz., significant factor loadings), but were different enough from each other to clearly represent separate dimensions (discriminant validity). Overall model fit statistics were all within acceptable limits.

RELIABILITY

Smith et al. assessed the extent to which the respondents were giving true scores by using both internal consistency measures (two forms of this were employed via the LISREL analysis) and test-retest. The instrument proved to be reliable by generally accepted internal consistency standards. The overall test-retest coefficient was .78.

NOMOLOGICAL VALIDITY

To examine whether the constructs of information privacy could find support in a network of theoretical relationships, Smith et al. looked at the linkages between standard antecedents of concerns of information privacy and items in the refined instrument. For this test, 77 business graduate students from two geographically dispersed U.S. universities completed the instrument, to which were added the questions measuring the exogenous variables. Significant beta coefficients in a regression analysis indicated that the instrument demonstrated nomological validity. Further tests of this validity were made through examining linkages to the personality characteristics of: (a) trust/distrust, (b) paranoia, and (c) social criticism. For these tests, a new sample of undergraduate students was used.

PREDICTIVE VALIDITY

The practical test of the instrument was to determine whether its measures would correlate highly with criteria in previous public opinion surveys administered for or by Cambridge Reports and Equifax. Three questions used on these surveys were included with the Smith et al. instrument, and the combined survey administered to 354 members of the Information Systems Audit and Control Association (ISACA). Highly significant correlations were observed between an overall index of Smith et al. items and the previous survey items.

EXTERNAL VALIDITY

Because consistent results were found across sample groups as diverse as IS auditors, knowledge workers in banking, insurance and credit, and graduate and undergraduate business students, the authors conclude that the instrument will generalize well.

V. CONCLUSION

In conclusion, we wish to stress that validation guidelines are owned by communities of practice and should not be the provenance of methodological “experts” or people in positions of power. Moreover, our argumentation in this paper should not be viewed as in any sense “conclusive.”

In fact, we very much welcome criticism of the logic we followed, the examples, empirical evidence, and authorities cited. If there are reasons why the IS positivist community should not view internal consistency reliability as a “must do” for researchers, then we need to hear these objections. Likewise for the other validities. This article is offered in the spirit of initiating a debate on the critical issue of what “rigor” in IS research means.

In that this research essay is and must be in the form of a philosophical disputation with support from the methodological literature, the heuristics presented here are clearly subject to debate. Notwithstanding their usefulness to guide research for the interim, the IS field should welcome an ongoing discussion of key methodological issues.[31] It is clear that a wide variety of methodology specialists within the IS field are capable of articulating the principles that guide their practice. To encourage this dialogue could, one might even argue “should”, be a worthy goal of IS journals. The quality of our science should be *sine qua non*, “without which nothing.”

Other fields looked at their research strategies and extent of validation introspectively (e.g., [Scandura and Williams, 2000]). Much of what we believe, for example, resulted from a series of books and articles by psychologists. These researchers, along with others, remind us that positivist science needs to be more than a series of anecdotes or highly biased observations. It needs the rigor of careful and thoughtful data gathering and intellectual constructs that explain real world events. Validating the quantitative, positivist approach that one takes so that other

[31] See http://www.endnote.auckland.ac.nz/ for a bibliography developed entirely to methodology and research validities. This bibliography is a starting point for more work in methodology.

scientists test or extend one's work is a critical underpinning of the scientific endeavor. Across-the-board validation of our research — regardless of choice of methodology — could be our next community goal. Heuristics and guidelines for bringing more rigor to the process of scientific investigation are offered in this paper. The gatekeepers of the field, as represented by the journals and conferences, need to raise the level of awareness of the entire community by insisting on the standards offered here or convincingly presented by others.

Editor's Note: This article was fully peer reviewed. It was received on February 27, 3002. It was with the authors 7 months for revisions. The article was published onApril 30, 2004.

REFERENCES

Adams, D. A., R. R. Nelson, and P. A. Todd (1992) "Perceived Usefulness, Ease of Use, and Usage of Information Technology: A Replication," *MIS Quarterly* (16) 2, June, pp. 227-248.

Alreck, P. L. and R. B. Settle (1995) "Planning Your Survey," *American Demographics* pp. 12.

Alwin, D. (1973-74) "Approaches to the Interpretation of Relationships in the Multitrait-Multimethod Matrix," *Sociological Methodology* pp. 79-105.

Anderson, J. C. and D. W. Gerbing (1988) "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin* (103) 3, Fall, pp. 411-423.

Anderson, J. C., D. W. Gerbing, and J. E. Hunter (1987) "On the Assessment of Unidimensional Measurement: Internal and External Consistency, and Overall Consistency Criteria," *Journal of Marketing Research* (24pp. 432-437.

Andrews, F. M. (1984) "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach," *Public Opinion Quarterly* (48) 2, Summer, pp. 409-442.

Argyris, C. (1979) "Some Unintended Consequences of Rigorous Research," in R. T. Mowday and R. M. Steers (Eds.) *Research in Organizations: Issues and Controversies*, Santa Monica, CA: Goodyear, pp. 290-304.

Armstrong, D., A. Gosling, J. Weinman, and T. Marteau (1997) "The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study," *Sociology: The Journal of the British Sociological Association* (31) 3, August, pp. 597-606.

Avolio, B. J., F. J. Yammarino, and B. M. Bass (1991) "Unresolved Sticky Issue," *Journal of Management* (17) 3, September, pp. 571-587.

Bagozzi, R. P. (1977) "Structural Equation Models in Experimental Research," *Journal of Marketing Research* (14pp. 209-236.

Bagozzi, R. P. (1980) *Causal Methods in Marketing*. New York: John Wiley and Sons.

Bagozzi, R. P. and C. Fornell (1982) "Theoretical Concepts, Measurement, and Meaning," in, vol. 2 C. Fornell (Ed.) *A Second Generation of Mulivariate Analysis*: Praeger, pp. 5-23.

Bagozzi, R. P. and L. W. Phillips (1982) "Representing and Testing Organizational Theories: A Holistic Construal," *Administrative Science Quarterly* (27) 3, pp. 459-489.

Bagozzi, R. P., Y. Yi, and L. W. Phillips (1991) "Assessing Construct Validity in Organizational Research," *Administrative Science Quarterly* (36) 3, September, pp. 421-458.

Barki, H. and J. Hartwick (1994) "User Participation, Conflict, and Conflict Resolution: The Mediating Roles of Influence," *Information Systems Research* (5) 4, December, pp. 422-438.

Barrett, R. S. (1980-81) "Is the Test Content-Valid: Or, Does It Really Measure a Construct?," *Employee Relations Law Journal* (6) 3, Winter, pp. 459-475.

Barrett, R. S. (1981) "Is the Test Content-Valid: Or, Who Killed Cock Robin?," *Employee Relations Law Journal* (6) 4, Spring, pp. 584-600.

Blalock, H. M. (1969) *Theory Construction: From Verbal to Mathematical Formulations*. Englewood Cliffs, NJ: Prentice-Hall.

Bottger, P. and P. Yetton (1982) "Student Assessment and GMAT: Quantitative Versus Verbal Components in Performance on Examinations and Assignments," *Australian Journal of Management* (7) 1, June, pp. 9-18.

Boudreau, M., D. Gefen, and D. Straub (2001) "Validation in IS Research: A State-of-the-Art Assessment," *MIS Quarterly* (25) 1, March, pp. 1-23.

Boudreau, M.-C., T. Ariyachandra, D. Gefen, and D. Straub (2004) "Validating IS Positivist Instrumentation: 1997-2001," in M. E. Whitman and A. B. Woszczynski (Eds.) *The Handbook of Information Systems Research*, Hershey, PA USA: Idea Group Publishing, pp. 15-26.

Bowers, J. W. and J. A. Courtright (1984) *Communication Research Methods*. Glenview, IL: Scott, Foresman.

Brock, D. B. and L. M. Sulsky (1994) "Attitudes toward Computers: Construct Validation and Relations to Computer Use," *Journal of Organizational Behavior* (15) 1, January, pp. 17-35.

Burrell, G. and G. Morgan (1979) *Sociological Paradigms and Organizational Analysis*. Portsmouth, NH: Heinemann.

Campbell, D. T. (1960) "Recommendations for APA Test Standards Regarding Construct, Trait, Discriminant Validity," *American Psychologist* (15) August, pp. 546-553.

Campbell, D. T. (1975) "Degrees of Freedom and the Case Study," *Comparative Political Studies* (8) 2, July, pp. 178-193.

Campbell, D. T. and D. W. Fiske (1959) "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin* (56) 2, March, pp. 81-105.

Carrier, M. R., A. T. Dalessio, and S. H. Brown (1990) "Correspondence Between Estimates of Content and Criterion-Related Validity Values," *Personnel Psychology* (43) 1 (Spring), pp. 85-100.

Chin, W. W. (1998a) "Issues and Opinion on Structural Equation Modeling," *MIS Quarterly* (22) 1, March, pp. vii-xvi.

Chin, W. W. (1998b) "The Partial Least Squares Approach to Structural Equation Modeling," in G. A. Marcoulides (Ed.) *Modern Methods for Business Research*, London, pp. 295-336.

Chin, W. W., A. Gopal, and W. D. Salisbury (1997) "Advancing the Theory of Adaptive Structuration: The Development of a Scale to Measure Faithfulness of Appropriation," *Information Systems Research* (8) 4, pp. 342-367.

Churchill, G. A., Jr. (1979) "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research* (16) 1, February, pp. 64-73.

Cohen, J. (1977) *Statistical Power Analysis for the Behavioral Sciences*, Revised Edition edition. New York: Academic Press.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Hillsdale, NJ: L. Erlbaum Associates.

Collopy, F., M. Adya, and J. S. Armstrong (1994) "Principles for Examining Predictive Validity: The Case of Information Systems Spending Forecasts," *Information Systems Research* (5) 2, June, pp. 170-179.

Cook, T. D. and D. T. Campbell (1979) *Quasi Experimentation: Design and Analytical Issues for Field Settings*. Chicago: Rand McNally.

Coombs, C. H. (1976) *A Theory of Data*. Ann Arbor, MI: Mathesis Press.

Cronbach, L. J. (1951) "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika* (16) September, pp. 297-334.

Cronbach, L. J. (1971) "Test Validation," in 2nd Edition edition R. L. Thorndike (Ed.) *Educational Measurement*, Washington, D.C.: American Council on Education, pp. 443-507.

Cronbach, L. J. (1990) *Essentials of Psychological Testing*, 5th edition. New York: Harper-Row.

Cronbach, L. J. and P. E. Meehl (1955) "Construct Validity in Psychological Tests," *Psychological Bulletin* (55) 4, July, pp. 281-302.

Davis, F. D. (1989) "Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology," *MIS Quarterly* (13) 3, September, pp. 319-340.

Denzin, N. K. and Y. S. Lincoln (1994) "Introduction: Entering the Field of Qualitative Research," in N. K. Denzin and Y. S. Lincoln (Eds.) *Handbook of Qualitative Research*, London: Sage, pp. 1-18.

Devaraj, S., M. Fan, and R. Kohli (2002) "Antecedents of B2C Channel Satisfaction and Preference: Validating e-commerce Metrics," *Information Systems Research* (13) 3, September, pp. 316-333.

Dholakia, U. M. and R. P. Bagozzi (2003) "As Time Goes By: How Goal and Implementation Intentions Influence Enactment of Short-Fuse Behaviors," *Journal of Applied Social Psychology* (33) 5, May, pp. 889-923.

Diamantopoulos, A. and H. M. Winklhofer (2001) "Index Construction with Formative Indicators: An Alternative to Scale Development," *Journal of Marketing Research* (38) 2, pp. 269-277.

Doll, W. J. and G. Torkzadeh (1988) "The Measurement of End-User Computing Satisfaction," *MIS Quarterly* (12) 2, pp. 259-274.

Falk, R. F. and N. B. Miller (1992) *A Primer for Soft Modeling*. Akron, OH: University of Akron Press.

Fornell, C. and D. Larcker (1981) "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18pp. 39-50.

Gefen, D. (1997) Building Users' Trust in Freeware Providers and the Effects of this Trust on Users' Perceptions of Usefulness, Ease of Use and Intended Use. Dissertation, Georgia State University.

Gefen, D. (2000) "It is Not Enough To Be Responsive: The Role of Cooperative Intentions in MRP II Adoption," *DATA BASE for Advances in Information System* (31) 2, pp. 65-79.

Gefen, D. (2003) "Unidimensional Validity: An Explanation and Example," *CAIS* (12) 2, pp. 23-47.

Gefen, D., E. Karahanna, and D. Straub (2003) "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27) 1, March, pp. 51-90.

Gefen, D., D. Straub, and M. Boudreau (2000) "Structural Equation Modeling Techniques and Regression: Guidelines for Research Practice," *Communications of AIS* (7) 7 August,, pp. 1-78.

Gefen, D. and D. W. Straub (1997) "Gender Differences in Perception and Adoption of E-Mail: An Extension to the Technology Acceptance Model," *MIS Quarterly* (21) 4 (December), pp. 389-400.

Gerbing, D. W. and J. C. Anderson (1988) "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," *Journal of Marketing Research* (25) May, pp. 186-192.

Grover, V., M. J. Cheon, and J. T. C. Teng (1996) "The Effect of Service Quality and Partnership on the Outsourcing of Information Systems Functions," *Journal of Management Information Systems* (12) 4, Spring, pp. 89-116.

Guion, R. M. (1977) "Content Validity: Three Years of Talk-What's the Action?," *Public Personnel Management* (6) 6 (November-December), pp. 407-414.

Hair, J. F., Jr., R. E. Anderson, R. L. Tatham, and W. C. Black (1998) *Multivariate Data Analysis with Readings, 5th Edition*. Englewood Cliffs, NJ: Prentice Hall.

Hambleton, R., H. Swaminathan, and H. Swaminathan (1984) *Item Response Theory: Principles and Applications*. Rotterdam: Kluwer Academic Publishers.

Hendrickson, A., K. Glorfeld, and T. P. Cronan (1994) "On the Repeated Test-Retest Reliability of the End-User Computing Satisfaction Instrument: A Comment," *Decision Sciences* (25) 4 (July-August), pp. 655-667.

Hendrickson, A. R., P. D. Massey, and T. P. Cronan (1993) "On the Test-Retest Reliability of Perceived Usefulness and Perceived Ease of Use Scales," *MIS Quarterly* (17) 2 (June), pp. 227-230.

Igbaria, M. and J. J. Baroudi (1993) "A Short-Form Measure of Career Orientations: A Psychometric Evaluation," *Journal of Management Information Systems* (10) 2, Fall, pp. 131-154.

Jarvenpaa, S. L., G. W. Dickson, and G. DeSanctis (1985) "Methodological Issues in Experimental IS Research: Experiences and Recommendations," *MIS Quarterly* (9) 2 (June), pp. 141-156.

Jenkins, A. M. (ed.) (1985) *Research Methodologies and MIS Research*. *Research Methods in Information Systems*, Amsterdam, Holland: Elsevier Science Publishers B.V.

Jones, A. P., L. A. Johnson, M. C. Butler, and D. S. Main (1983) "Apples and Oranges: An Empirical Comparison of Commonly Used Indices of Interrater Agreement," *Academy of Management Journal* (26) 3, September, pp. 507-519.

Jöreskog, K. G. and D. Sörbom (1989) *LISREL7: A Guide to the Program and Applications*, 2nd edition. Chicago: SPSS Inc.

Jöreskog, K. G. and D. Sörbom (1993) *LISREL8: Structural Equation Modeling with SIMPLIS Command Language*, 2nd edition. Chicago, IL: Scientific Software International.

Jöreskog, K. G. and D. Sörbom (1994) *LISREL 8.12i and PRELIS 2.12i for Windows*. Chicago: Scientific Software International.

Karahanna, E., D. W. Straub, and N. L. Chervany (1999) "Information Technology Adoption across Time: A Cross-Sectional Comparison of Pre-Adoption and Post-Adoption Beliefs," *MIS Quarterly* (23) 2, pp. 183-213.

Keil, M., D. P. Truex, and R. Mixon (1995) "The Effects of Sunk Cost and Project Completion on Information Technology Project Escalation," *IEEE Transactions on Engineering Management* (42) 4, November, pp. 372-381.

Kerlinger, F. N. (1964) *Foundations of Behavioral Research*. New York: Holt, Rinehart, and Winston.

Kuhn, T. S. (1970) *The Structure of Scientific Revolutions*, 2nd edition. Chicago, IL: The University of Chicago Press.

Lacity, M. C. and M. A. Janson (1994) "Understanding Qualitative Data: A Framework of Text Analysis Methods," *Journal of Management Information Systems: JMIS* (11) 2 (Fall), pp. 137-155.

Landis, J. R. and G. G. Koch (1977) "The Measurement of Observer Agreement for Categorical Data," *Biometrics* (22pp. 79-94.

Lawshe, C. H. (1975) "A Quantitative Approach to Content Validity," *Personnel Psychology* (28pp. 563-575.

Lawther, W. C. (1986) "Content Validation: Conceptual and Methodological Issues," *Review of Public Personnel Administration* (6) 3 (Summer), pp. 37-49.

Lewis, B. R., C. A. Snyder, and R. K. Rainer, Jr. (1995) "An Empirical Assessment of the Information Resource Management Construct," *Journal of Management Information Systems* (12) 1, Summer, pp. 199-223.

Lim, K. H., L. M. Ward, and I. Benbasat (1997) "An Empirical Study of Computer System Learning: Comparison of Co-Discovery and Self-Discovery Methods," *Information Systems Research* (8) 3, September, pp. 254-272.

Loch, K., D. Straub, and S. Kamel (2003) "Diffusing the Internet in the Arab World: The Role of Social Norms and Technological Culturation," *IEEE Transactions on Engineering Management* (50) 1, February, pp. 45-63.

Long, J. S. (1983a) *Confirmatory Factor Analysis: A Preface to LISREL*. Beverley Hills, CA USA: Sage.

Long, J. S. (1983b) *Covariance Structure Models : An Introduction to LISREL*. Beverly Hills, CA: Sage.

MacCallum, R. C. and J. T. Austin (2000) "Applications of Structural Equation Modeling in Psychological Research," *Annual Review of Psychology* (51pp. 201-226.

Marks, R., P. Watt, and P. Yetton (1981) "GMAT Scores and Performance: Selecting Students into a Graduate Management School," *Australian Journal of Management* (6) 2 (December), pp. 81-102.

Marsh, H. W. and D. Hocevar (1988) "A New, Powerful Approach to Multitrait-Multimethod Analyses: Application of Second Order Confirmatory Factor Analysis," *Journal of Applied Psychology* (73) 1, pp. 107-117.

Massetti, B. (1996) "An Empirical Examination of the Value of Creativity Support Systems on Idea Generation," *MIS Quarterly* (20) 1, March, pp. 83-97.

McKnight, D. H., V. Choudhury, and C. Kacmar (2002a) "Developing and Validating Trust Measures for E-Commerce: An Integrative Typology," *Information Systems Research* (13) 3, September, pp. 334-359.

McKnight, H., V. Choudhury, and C. Kacmar (2002b) "Developing and Validating Trust Measures for E-Commerce: An Integrative Typology," *Information Systems Research* (13) 3, pp. 334-359.

McLean, E. R., S. J. Smits, and J. R. Tanner (1996) "The Importance of Salary on Job and Career Attitudes of Information Systems Professionals," *Information & Management* (30) 6 (September), pp. 291-299.

Meehl, P. E. (1967) "Theory-Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*) June, pp. 103-115.

Miles, M. B. and A. M. Huberman (1994) *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: Sage Publications, Inc.

Millsap, R. E. (1990) "A Cautionary Note on the Detection of Method Variance in Multitrait-Multimethod Data," *Journal of Applied Psychology* (75) 3 (June), pp. 350-353.

Moore, G. C. and I. Benbasat (1991) "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation," *Information Systems Research* (2) 3 (September), pp. 192-222.

Mumford, M. D. and G. S. Stokes (1992) “Developmental Determinants of Individual Action: Theory and Practice in the Application of Background Data Measures,” in, vol. 2 M. D. Dunnette and L. M. Hough (Eds.) *Handbook of Industrial and Organizational Psychology*, Palo Alto, CA USA: Consulting Psychologists Press, pp. 61-138.

Netemeyer, R. G., S. Durvasula, and D. R. Lichtenstein (1991) "A Cross-National Assessment of the Reliability and Validity of the CETSCALE," *Journal of Marketing Research* (28) 3, pp. 320-327.

Netemeyer, R. G., D. A. Willamson, S. Burton, D. Biswas et al. (2002) "Psychometric Properties of Shortened Versions of the Automatic Thoughts Questionnaire," *Educational & Psychological Measurement* (62) 1, February, pp. 111-130.

Nielsen, I. K., S. M. Jex, and G. A. Adams (2000) "Development and Validation of Scores on a Two-Dimensional Workplace," *Educational & Psychological Measurement* (60) 4, August, pp. 628-644.

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

Nunnally, J. C. (1978) *Psychometric Theory*, 2nd edition. New York: McGraw-Hill.

Nunnally, J. C. and I. H. Bernstein (1994) *Psychometric Theory*, 3rd edition. New York: McGraw-Hill.

Orne, M. T. (1962) "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist* (17) 11, pp. 776-783.

Orne, M. T. (1969) “Demand Characteristics and the Concept of Quasi-Controls,” in R. Rosenthal and R. L. Rosnow (Eds.) *Artifact in Behavioral Research*, New York: Academic Press, pp. 143-179.

Parameswaran, R., B. A. Greenberg, D. N. Bellenger, and D. H. Roberson (1979) "Measuring Reliability: A Comparison of Alternative Techniques," *Journal of Marketing Research* (16) 1, February, pp. 18-25.

Perdue, B. C. and J. O. Summers (1986) "Checking the Success of Manipulations in Marketing Experiments," *Journal of Marketing Research* (23) 4 (November), pp. 317-326.

Perreault, W. D., Jr. and L. E. Leigh (1989) "Reliability of Nominal Data Based on Qualitative Judgments," *Journal of Marketing Research* (26) 2 (May), pp. 135-148.

Peter, J. P. (1979) "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," *Journal of Marketing Research* (16) 1 (February), pp. 6-17.

Pinsonneault, A. and N. Heppel (1997/98) "Anonymity in Group Support Systems Research: A New Conceptualization, Measure, and Contingency Framework," *Journal of Management Information Systems* (14) 3, Winter, pp. 89-108.

Pitt, L. F., R. T. Watson, and C. B. Kavan (1995) "Service Quality: A Measure of Information Systems Effectiveness," *MIS Quarterly* (19) 2, June, pp. 173-187.

Podsakoff, P. M., S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff (2003) "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology* (88) 5, pp. 879-903.

Ravichandran, T. and A. Rai (2000) "Quality Management in Systems Development: An Organizational System Perspective," *MIS Quarterly* (24) 3, September, pp. 381-415.

Rogers, T. B. (1995) *The Psychological Testing Enterprise*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Sambamurthy, V. and W. W. Chin (1994) "The Effects of Group Attitudes toward Alternative GDSS Designs on the Decision-making Performance of Computer-Supported Groups," *Decision Science* (25) 2, pp. 215-239.

Scandura, T. A. and E. A. Williams (2000) "Research Methodology in Management: Current Practices, Trends, and Implications for Future Research," *Academy of Management Journal* (43) 6, pp. 1248-1264.

Segars, A. H. (1997) "Assessing the Unidimensionality of Measurement: A Paradigm and Illustration within the Context of Information Systems Research," *OMEGA* (25) 1, February, pp. 107-121.

Segars, A. H. and V. Grover (1998) "Strategic Information Systems Planning Success: An Investigation of the Construct and its Measurement," *MIS Quarterly* (22) 2, June, pp. 139-163.

Sethi, V. and W. R. King (1994) "Development of Measures to Assess the Extent to Which an Information Technology Application Provides Competitive Advantage," *Management Science* (40) 12, December, pp. 1601-1627.

Simon, H. (1981) *The Sciences of the Artificial*, 2nd edition. Cambridge, MA: MIT Press.

Simon, S. J., V. Grover, J. T. C. Teng, and K. Whitcomb (1996) "The Relationship of Information System Training Methods and Cognitive Ability to End-user Satisfaction, Comprehension, and Skill Transfer: A Longitudinal Field Study," *Information Systems Research* (7) 4, pp. 466-490.

Smith, H. J., S. J. Milberg, and S. J. Burke (1996) "Information Privacy: Measuring Individuals' Concerns about Organizational Practices," *MIS Quarterly* (20) 2, June, pp. 167-196.

Spector, P. E. (1987) "Method Variance as an Artifact in Self-Reported Affect and Perceptions at Work: Myth or Significant Problem?," *Journal of Applied Psychology* (72) 3 (August), pp. 438-443.

Storey, V., D. Straub, K. Stewart, and R. Welke (2000) "A Conceptual Investigation of the Electronic Commerce Industry," *Communications of the ACM* (43) 7 (July), pp. 117-123.

Straub, D. W. (1989) "Validating Instruments in MIS Research," *MIS Quarterly* (13) 2, pp. 147-169.

Straub, D. W. (1990) "Effective IS Security: An Empirical Study," *Information Systems Research* (1) 3, pp. 255-276.

Straub, D. W. (1994) "The Effect of Culture on IT Diffusion: E-Mail and FAX in Japan and the U.S.," *Information Systems Research* (5) 1, March, pp. 23-47.

Straub, D. W. and E. Karahanna (1998) "Knowledge Worker Communications and Recipient Availability: Toward a Task Closure Explanation of Media Choice," *Organization Science* (9) 2 (March), pp. 160-175.

Straub, D. W., M. Limayem, and E. Karahanna (1995) "Measuring System Usage: Implications for IS Theory Testing," *Management Science* (41) 8, August, pp. 1328-1342.

Sussmann, M. and D. U. Robertson (1986) "The Validity of Validity: An Analysis of Validation Study Designs," *Journal of Applied Psychology* (71) 3, pp. 461-468.

Szajna, B. (1994) "Software Evaluation and Choice: Predictive Validation of the Technology Acceptance Instrument," *MIS Quarterly* (17) 3, pp. 319-324.

Taylor, S. and P. A. Todd (1995) "Understanding Information Technology Usage: A Test of Competing Models," *Information Systems Research* (6) 2, June, pp. 144-176.

Thomas, D. M. and R. T. Watson (2002) "Q-sorting and MIS Research: A Primer," *CAIS* (8) Article 9, pp. 141-156.

Thompson, R., D. W. Barclay, and C. A. Higgins (1995) "The Partial Least Squares Approach to Causal Modeling: Personal Computer Adoption and Use as an Illustration," *Technology Studies: Special Issue on Research Methodology* (2) 2, Fall, pp. 284-324.

Torkzadeh, G. and W. J. Doll (1994) "The Test-Retest Reliability of User Involvement Instruments," *Information & Management* (26) 1, pp. 21-31.

Umesh, U. N., R. A. Peterson, and M. H. Sauber (1989) "Interjudge Agreement and the Maximum Value of Kappa," *Educational and Psychological Measurement* (49pp. 835-850.

Van Dyke, T. P., L. A. Kappelman, and V. R. Prybutok (1997) "Measuring Information Systems Service Quality: Concerns on the Use of the SERVQUAL Questionnaire," *MIS Quarterly* (21) 2, June, pp. 195-208.

Venkatraman, N. and V. Ramanujam (1987) "Measurement of Business Economic Performance: An Examination of Method Convergence," *Journal of Management* (13) 1, Spring, pp. 109-122.

Webster, J. and D. Compeau (1996) "Computer-Assisted versus Paper-and-Pencil Administration of Questionnaires," *Behavior Research Methods, Instruments, & Computers* (28) 4, November, pp. 567-577.

Westland, J. C. (2004) "The IS Core XII: Authority, Dogma, and Positive Science in Information Systems Research," *CAIS* (13) 12, February, pp. 136-157.

Woszczynski, A. B. and M. E. Whitman (2004) "The Problem of Common Method Variance in IS Research," in M. E. Whitman and A. B. Woszczynski (Eds.) *The Handbook of Information Systems Research*, Hershey, PA USA: Idea Group Publishing, pp. 66-77.

GLOSSARY

AGFI: Adjusted Goodness of Fit Index. Within covariance-based SEM, statistic measuring the fit (adjusted for degrees of freedom) of the combined measurement and structural model to the data.

AMOS: A covariance-based SEM, developed by Dr. Arbuckle, Published by SmallWaters and marketed by SPSS as a statistically equivalent tool to LISREL. Details are available at http://www.spss.com/amos/ .

ANOVA: Univariate analysis of variance. Statistical technique to determine, on the basis of one dependent measure, whether samples are from populations with equal means.

AVE: Average Variance Extracted. Calculated as $[(\Sigma\lambda_i^2)/((\Sigma\lambda_i^2) + (\Sigma(1-\lambda_i^2)]$, the AVE measures the percent of variance captured by a construct by showing the ratio of the sum of the variance captured by the construct and measurement variance.

CFA: Confirmatory Factor Analysis. A variant of factor analysis where the goal is to test specific theoretical expectations about the structure of a set of measures.

Construct validity: One of a number of subtypes of validity that focuses on the extent to which a given test/instrumentation is an effective measure of a theoretical construct.

Content validity: The degree to which items in an instrument reflect the content universe to which the instrument will be generalized. This validity is generally established through literature reviews and expert judges or panels.

Cronbach's α: Commonly used measure of reliability for a set of two or more construct indicators. Values range between 0 and 1.0, with higher values indicating higher reliability among the indicators.

Dependent Variable (DV): Presumed effect of, or response to, a change in the independent variable(s).

EQS: A covariance-based SEM developed by Dr. Bentler and sold by Multivariate Software, Inc. EQS provides researchers with the ability to perform a wide array of analyses, including linear regressions, CFA, path analysis, and population comparisons. Details are available at http://www.smallwaters.com/.

Endogenous construct: Construct that is the dependent or outcome variable in at least one causal relationship. In terms of a path diagram, there are one or more arrows leading into the endogenous construct.

Exogenous construct: Construct that acts only as a predictor or "cause" for other constructs in the model. In terms of a path diagram, the exogenous constructs have only causal arrows leading out of them and are not predicted by any other constructs in the model.

Factor analysis: A statistical approach that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors).

Formative variables: Observed variables that “cause” the latent variable, i.e., represent different dimensions of it.

GFI: Goodness of Fit Index. Within covariance-based SEM, statistic measuring the absolute fit (unadjusted for degrees of freedom) of the combined measurement and structural model to the data.

Independent Variable (IV): Presumed cause of any change in a response or dependent variable(s).

Latent variable or construct: Research construct that is not observable or measured directly, but is measured indirectly through observable variables that reflect or form the construct.

Linear regression: A linear regression uses the method of least squares to determine the best equation describing a set of x and y data points.

LISREL: A procedure for the analysis of **LI**near **S**tructural **REL**ations among one or more sets of variables and variates. It examines the covariance structures of the variables and variates included in the model under consideration. LISREL permits both confirmatory factory analysis and the analysis of path models with multiple sets of data in a simultaneous analysis.

Loading (Factor Loading): Weighting which reflect the correlation between the original variables and derived factors. Squared factor loadings are the percent of variance in an observed item that is explained by its factor.

Manipulation validity: A measure of the extent to which treatments have been perceived by the subjects of an experiment.

Measurement model: Sub-model in structural equation modeling that (1) specifies the indicators for each construct, and (2) assesses the reliability of each construct for estimating the causal relationships.

MTMM: Multitrait-multimethod matrices employ correlations representing all possible relationships between a set of constructs, each measured by the same set of methods. This matrix is one of many methods that can be used to evaluate construct validity by demonstrating both convergent and discriminant validity.

NFI: Normed Fix Index. Within covariance-based SEM, statistic measuring the normed difference in χ^2 between a single factor null model and a proposed multi-factor model.

Observed indicator / variables: Observed value used as an indirect measure of a concept or latent variable that cannot be measured or observed directly.

Parallel correlational patterns (see Unidimensionality**):** Additional correlations between measurement items that are not reflected in a factor analysis or in the measurement model. For example, if items A1, A2, A3 and A4 load together on the same factor in a factor analysis but, additionally, A1 and A2 are highly correlated to each other in another dimension that is not captured in the factor analysis. Confirmatory factor analysis in LISREL can detect such cases.

PLS: Partial Least Squares. A second generation regression model that combines a factor analysis with linear regressions, making only minimal distribution assumptions.

PCA: Principal Components Analysis. Statistical procedure employed to resolve a set of correlated variables into a smaller group of uncorrelated or orthogonal factors.

Q-sort: A modified rank-ordering procedure in which stimuli are placed in an order that is significant from the standpoint of a person operating under specified conditions. It results in the captured patterns of respondents to the stimulus presented. Those patterns can then be analyzed to discover groupings of response patterns, supporting effective inductive reasoning.

Reflective variables: Observed variables that "reflect" the latent variable and as a representation of the latent variable should be unidimensional and correlated.

Reliability: Extent to which a variable or set of variables is consistent in what it is intended to measure. If multiple measurements are taken, the reliable measures will all be very consistent in their values. Reliable measures approach a true, but unknown "score" of a construct.

R-square or R^2: Coefficient of determination. Measure of the proportion of the variance of the dependent variable about its mean that is explained by the independent variable(s). R-square is derived from the F statistic. This statistic is usually employed in linear regression analysis and PLS.

SEM: Structural Equation Modeling. Multivariate technique combining aspects of multiple regression (examining dependence relationships) and factor analysis (representing unmeasured concepts with multiple variables) to estimate a series of interrelated dependence relationships simultaneously.

Statistical conclusion validity: Type of validity that addresses whether appropriate statistics were used in calculations that were performed to draw conclusions about the population of interest.

Structural model: Set of one or more dependence relationships linking the model constructs. The structural model is most useful in representing the interrelationships of variables between dependence relationships.

Unidimensionality: A fundamental attribute of measurement items, assumed *a-priori* by scale reliability statistics. Unidimensional items occur when the items reflect only one underlying trait or concept. If a construct is unidimensional, a first order latent construct representing that variable will be superior to a set of second order constructs representing different aspects of a construct.

ABOUT THE AUTHORS

Marie-Claude Boudreau is Assistant Professor of MIS at the University of Georgia. She received her Ph.D. degree in Computer Information Systems from Georgia State University, a Diplôme d'Enseignement Supérieur Spécialisé from l'École Supérieure des Affaires de Grenoble, and an M.B.A. from l'Université Laval in Québec. Dr. Boudreau performs research on the implementation of integrated software packages and the organizational change induced by information technology. She is the author of articles published in such journals, as *Information Systems Research, MIS Quarterly*, *Journal of Management Information Systems, The Academy of Management Executive, CAIS,* and *Information Technology & People*

David Gefen is Associate Professor of MIS at Drexel, where he teaches Strategic Management of IT, Database, and VB.NET. He received his Ph.D. from Georgia State University and a Masters from Tel-Aviv University. His research focuses on psychological and rational processes in ERP, CMC, and e-commerce implementation. David's interests stem from 12 years developing and managing large IT systems. His research findings are published in *MISQ, ISR, IEEE TEM, JMIS, JSIS, DATA BASE, Omega, JAIS, CAIS,* and *JEUC.*

Detmar Straub is the J. Mack Robinson Distinguished Professor of Information Systems at Georgia State University and Vice President, Publications of AIS. Detmar researches net-

enhanced organizations and e-Commerce, computer security, technological innovation, and international IT. He holds a DBA (MIS; Indiana) and a PhD (English; Penn State). He published over 110 papers in journals such as *CAIS*, *JAIS*, *Management Science*, *Information Systems Research*, *MIS Quarterly*, *Organization Science*, *CACM*, *JMIS*, *Journal of Global Information Management*, *Information & Management*, *Academy of Management Executive*, and *Sloan Management Review*. He is currently a Senior Editor of *JAIS* and *DATA BASE* and an Associate Editor of *Management Science*. Former Co-Editor of *DATA BASE for Advances in Information Systems* and an Associate Editor and Associate Publisher for *MIS Quarterly*, he consulted widely in industry on computer security and technological innovation.