Validation Guidelines for IS Positivist Research

Detmar Straub* Marie-Claude Boudreau David Gefen*****

***Georgia State University**

Department of Computer Information Systems
College of Business Administration
Georgia State University
Atlanta, Georgia 30302-4015
dstraub@gsu.edu
404-651-3827 (P); 404-651-3842 (Fax)

****University of Georgia**

Management Information System Department
Terry College of Business
University of Georgia
Athens, Georgia 30602
mcboudre@terry.uga.edu

*****Drexel University**

Management Department
College of Business and Administration
Drexel University
32nd and Chestnut Streets
Philadelphia, PA 19104
gefend@drexel.edu

METHODOLOGY

Working paper version eventually to be published as: Straub, D.W., Boudreau, M.-C., and Gefen, D. "Validation Guidelines for IS Positivist Research," *CAIS* (forthcoming) 2004.

Copyright (c) 2004 by Detmar Straub, Marie-Claude Boudreau, and David Gefen
All rights reserved.

Validation Guidelines for IS Positivist Research

METHODOLOGY

Abstract

Fourteen years ago, Straub (1989) raised the issue of whether IS positivist researchers were sufficiently validating their instruments. Since then, the IS profession has been exposed to many opportunities and challenges. E-Commerce has risen in prominence. Other management trends have passed through their cycles. Our new professional society, the Association of Information Systems, has been formed and amalgamated with the preeminent research conference in the field, the International Conference on Information Systems.

But the issue of rigor in IS research is still one of our most critical scientific issues. Without solid validation of the instruments that are used to gather data upon which findings and interpretations are based, the very scientific basis of the profession is threatened.

This study builds on two prior retrospectives of IS research that conclude that IS positivist researchers still have major barriers to overcome 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, and this article is an initial 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 conclude the paper.

Keywords: IS research methods; measurement; psychometrics; validation; reliability; content validity; construct validity; statistical conclusion validity; nomological validity; predictive validity; concurrent validity; unidimensional reliability; factorial validity; manipulation

validity; formative measures; quantitative, positivist research; heuristics; guidelines; structural equation modeling; LISREL; PLS.

1. INTRODUCTION

Fourteen 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 has been exposed to many opportunities and challenges. E-Commerce has risen and, at least the term itself, has fallen in prominence. Other management trends have passed through their cycles. Our professional society, the Association of Information Systems, has been formed and amalgamated with the preeminent research conference in the field, the International Conference on Information Systems, and 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 had reported reliability of their scales, only 13% validated their constructs, while a scant 19% utilized 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. Two follow-up studies by 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 are simply not taking a stand at all on this issue. The paper is addressed at positivist researchers who already accept the epistemological stance that this 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 positivist intellectual position. Stated in absolute terms, this epistemological stance is that the world of phenomena has 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. Whereas this simple asseveration 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 articulated by positivist researchers are social constructions, albeit with "permanent" presence in the real world that allows them to be evaluated along the same lines as harder and less demonstrably subjective realities. In short, the willingness of contemporary positivist researchers to consider constructs as a "fuzzy" set rather than as the near perfect surrogate of an objective reality is sizeable.

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 constructions to be sure. Capturing these entities is a process that can be fraught with difficulties and one of the most tenacious of these is the inability of the IS community to know whether the measures being selected and utilized by the researchers are valid. This 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 these two recent analyses of IS research which, in brief, conclude that IS positivist researchers still have major barriers to overcome in instrument, statistical, and other forms of validation. This study goes far beyond Straub (1989), Gefen et al. (2000), and Boudreau et al. (2001) 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. 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. But 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.

In order to build our case, it is necessary to first discuss each validity at some length, which occurs in Section 2. 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,[1] which is found in Section 3. To provide extra guidance, each heuristic is qualified as being **mandatory**, **highly recommended**, **recommended**, or **optional.** Then, to demonstrate that these heuristics are attainable, an exemplar of how instruments can be developed is presented in Section 4. The final section offers concluding remarks.

2. 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 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 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 it is not directly observable. In short, there are no immediate and an obvious measure that the scientific community would agree on that captures 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 very pragmatic, concrete level. This paper attempts to address this issue by setting forth specific guidelines for validation, i.e., heuristics, 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

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

widely. Ironically, though, there are no simple answers to this question since scientific methods and techniques cannot themselves be used to validate the principles upon which they are based. *If that were so, the system would be purely tautological and reifying.* 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 in the present case, depend strictly on calls to authority and the persuasiveness of the ideas themselves. There are no established scientific standards against which to test them.

2.1 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). They are: (1) validation of data gathering, (2) ruling out rival hypotheses, (3) statistical inference, and (4) generalizability. Terms in widespread use for these validities are, respectively: (1) instrument / instrumentation validity, (2) internal validity, (3) statistical conclusion validity, and (4) 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.

To articulate the specific validation principles being referred to, we next briefly review and reassess those validation principles discussed in Straub (1989), both for the sake of completeness and in order to put addenda and new thoughts on IS instrument validities in the context of heuristics for practice. After this review and building of the cases for heuristics, findings from empirical studies of the use of validities in IS will also be more

meaningful. Validities to be reviewed along with heuristics and typical techniques are discussed with exemplars in IS research in Tables 1a and 1b below.

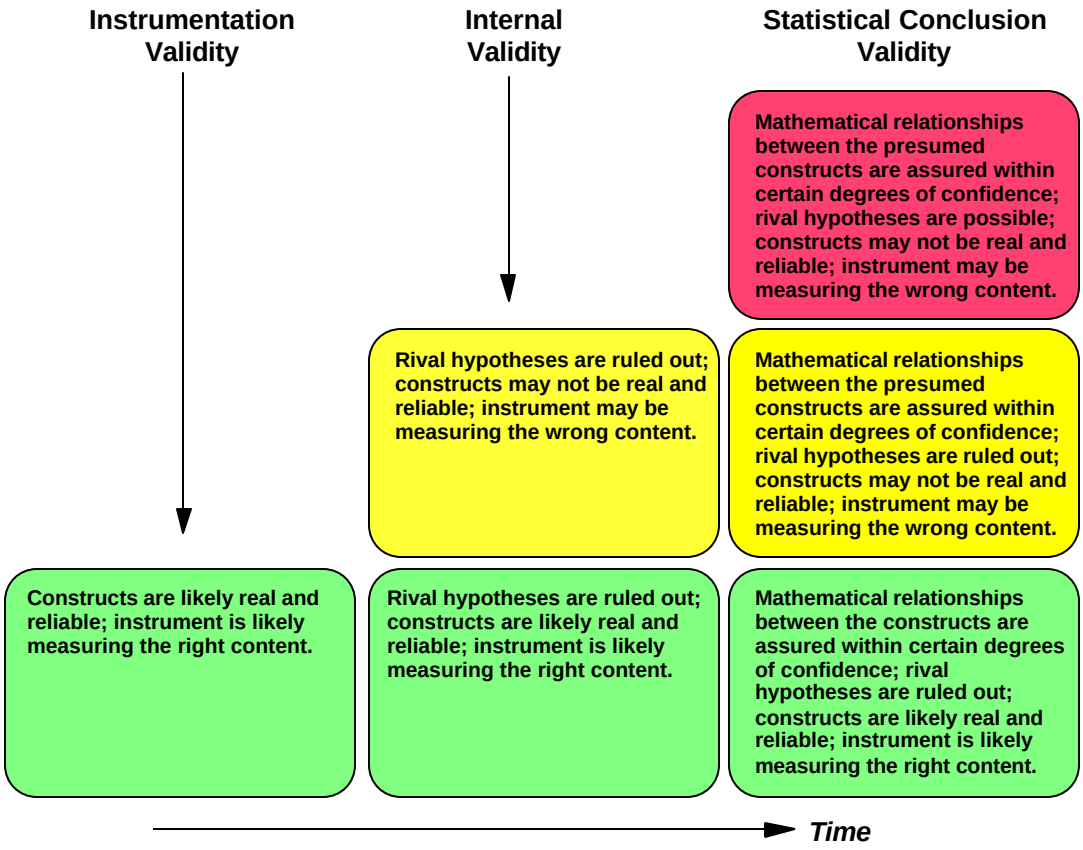


Figure 1. Validity Touchstones (Altered Somewhat from Straub, 1989)

(Legend: Green is preferred path; yellow is cautionary; red is least desirable path)

2.2 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)? As Figure 2 shows, researchers have many choices in creating means of measuring a construct. Have they chosen 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 well represent the construct, there will be measurement error. If they omit measures, the error is one of exclusion.

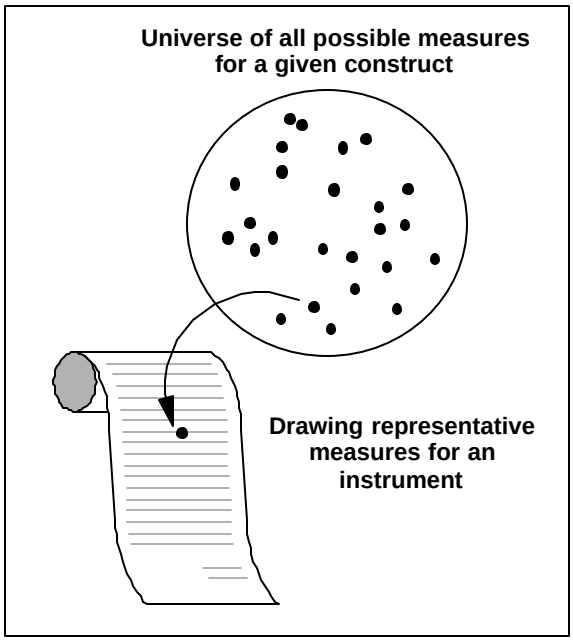


Figure 2. Pictorial Model of Content Validity

Whereas many psychometricians (Nunnally, 1978) and researchers (Barrett, 1980-81; Barrett, 1981) 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 a sampling issue and not an instrument evaluation issue (Guion, 1977).

Validity ✍✍ 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	Rare in IS research, MTMM has no well accepted statistical thresholds, 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, MTMM has no well accepted statistical thresholds, 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

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

Validity ✍✍ Validity Component	**Heuristics/ Techniques**	**Comments/ Pros and Cons**	**Examples in IS Research**
Reliability ✍✍Internal consistency	Cronbach's ?; correlations; SEM composite consistency estimates	alpha 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
✍✍Split half	Cronbach's ?; 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 ?	McLean et al., 1996
✍✍Test-retest	Cronbach's ?; correlations; SEM estimates	Comparisons across time of an instrument	Hendrickson et al., 1993 Torkzadeh and Doll, 1994
✍✍Alternative or equivalent forms	Cronbach's ?; 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 have been agreed upon; practice varies significantly.	Keil et al., 1995 Straub and Karahanna, 1998

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

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), there is no guarantee that the instrument items are randomly drawn in that the universe of these items itself are so indeterminate (Lawther, 1986; Nunnally, 1978). The most commonly employed evaluation of this validity is judgmental and highly subjective.[2] Moreover, it may well be, as Guion (1977) asserts, that content validity is really 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 the field of IS, Boudreau et al. (2001)[4] indicate that only 23% of the articles they sampled used 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 (1995) work. These authors validated the content of their information resource management (IRM) instrument via Lawshe's (1975) quantitative approach. 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 Lawshe's (1975) formulation.

[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.

Based on a table in Lawshe (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.

2.2..1 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. But at this point in the history of the positivist sciences, there is not a clear consensus on the methods and means of determining content validity. For this reason, we would argue that it is a **highly recommended**, but not mandatory practice for IS researchers.

2.3 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, 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 is a matter for content validity and straightforward definitions of the construct.

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 load on or are strongly associated with the blue construct, then they "converge" on this construct (convergent validity). If theoretically unrelated measures and constructs are entered, 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). Please 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 very robust one indeed, and it may not always work out. 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.

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 on its likeness or lack of similarity to strength of construct linkages in the past literature.

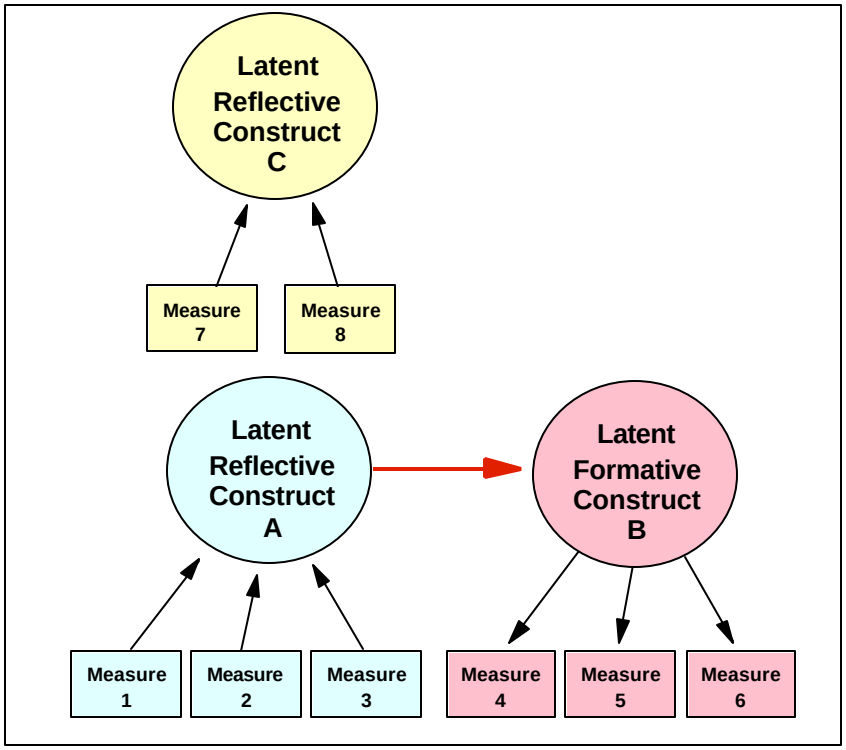


Figure 3. Pictorial Model of Construct Validity

2.3.1 Differences from Internal Validity and Forms of Construct Validity

As pointed out by Straub (1989), 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. 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 components of construct validity (Bagozzi, 1980). 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[5]. 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.

2.3.2 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.[6] Campbell and Fiske's (1959) multitrait-multimethod analysis (MTMM) 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.

Let us illustrate this with 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. So questions about perceived usefulness and ease of use are followed hard upon by questions about intention to use a system or its actual usage. This means of testing TAM has 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

[5] 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).

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

use or how useful a system is. Cognitive dissonance theory would suggest that respondents who felt that the system 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 formally test an instrument for common methods bias, two methods (i.e., instruments or data gathering-coding methods) are required (Straub, 1989). These 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 method 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. 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 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 was valid.

Other techniques besides these 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" (discussed in sub-section 2.3.3). But one such innovative means of verifying discriminant validity is Q-sorting (Moore and Benbasat, 1991; Segars and Grover, 1998; Storey et al., 2000). 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 et al., 2003, for a mainstream publication that assesses this.)

Vignette #2: Examples of Discriminant Validity

There are several examples in IS research other than Straub (1989) and Straub (1990) where a formal MTMM analysis was utilized 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 revealed 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 exemplar of a study using a different technique from MTMM to assess discriminant validity is Segars and Grover (1998). Here they use Q-sorting to validate the construct and sub-constructs of strategic information systems planning (SISP). After their literature review, four dimensions of SISP with 28 associated planning objectives were uncovered. A random listing of the 28 objectives in single sentence format were provided on pages separate from the 4 sub-construct dimensions of: (a) alignment, (b) analysis, (c) cooperation, and (d) 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.

2.3.3 Convergent Validity

Convergent validity is evidenced when items thought to reflect a construct converge, or show significant, high correlations with each other, particularly when compared to the convergence of items relevant to other constructs, irrespective of method.[7]

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

[7] 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 correlate (Jöreskog and Sörbom, 1989). See Gefen et al. (2000) for a review of this topic.

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 has 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).

As with discriminant validity, MTMM analysis of convergent validity is infrequent in IS research. MTMM demands the gathering of data through at least two "maximally different methods" (Campbell and Fiske, 1959, p. 83). This requirement 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 quite 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 will be delineated below, but IS researchers probably only need to apply MTMM after a research stream has matured (Straub, 1989) and there is a more exacting need to rule out methods bias. Clearly, there is a pressing need to scrutinize the entire TAM research stream for common methods bias; see Straub et al. (1995), for example.

In fact, there are numerous techniques available besides MTMM to evaluate convergent validity. In that these techniques are all based on variants of factor analysis, they will be discussed immediately below under "factorial validity."

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 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.

2.3.4 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,[8] which is by far the most frequent practice in IS research. Moreover, if two or more methods have been used to assess the instrument in question,[9] there is evidence that MTMM

[8] There is a continuing debate, it should be noted, as to whether methods bias is a significant problem in organizational research (Spector, 1987).

[9] 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

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).[10]

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 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.[11]

It is also important to recognize that the point of factorial validity is to examine how variables in each distinct causal stages 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 instance, 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*[12] 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.

---

these methods may be intended for a major data-gathering effort, as in Straub (1989), and, for that reason, we focus on validating a single instrument in the present study.

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

[11] There is some debate, however, about how to handle items that do not load properly. Churchill (1979) and Gerbing and Anderson (1988) suggest dropping such items, but since doing may result in over-fitting the model to the data, it is recommended in such cases to collect a second dataset and replicate the analysis on the reduced scales. There is also some debate about 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 et al., 1992). The important point to note is that factor analysis can show the way to "clean up" the construct.

[12] 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).

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, neglecting to explicitly specify that construct A is correlated with construct D when the items in construct A are highly correlated with those in construct D will result in unacceptably low fit-indexes (see Gefen et al., 2000, for a detailed discussion and comparison of covariance-based SEM, PLS, and linear regression).

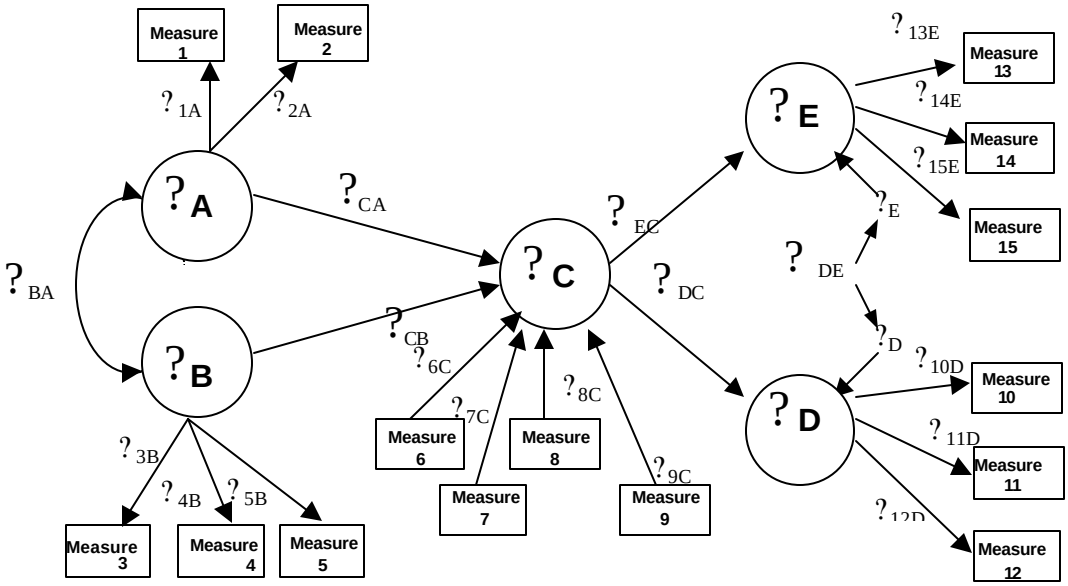


Figure 4. Generic Theoretical Network with Constructs and Measures (Adapted from Gefen et al., 2000)

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 1-5 (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.[13] 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.

[13] Control variables are often useful in running these tests.

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 be avoided in PCA.

SEM, on the other hand, facilitates the examination of factorial validity through a Confirmatory Factor Analysis (CFA). That is, by 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). See Gefen et al. (2000) for a detailed discussion.[14]

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, PLA, 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

[14] 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.

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.

2.3.5 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. As described in 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 have chosen 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 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. (2002) examines the psychometric properties of a trust instrument. To prove that trust is a multi-dimensional concept, they 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.

2.3.6 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: (a) the hypothesized relationships between the measurement items and their respective latent variables or (b) 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). 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 one concludes 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 have randomized 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 inherent constructs in the ordering of questionnaire items and are responding accordingly.[15] In Table 2, below, each column represents a possible ordering of questionnaire items of the original Davis (1989) instrument. 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 1-5 in column 1 are related (usage measures) as are items 6-11 (perceived usefulness measures) and items 12-17 (perceived ease-of-use measures). In short, the method itself is likely contributing to the pattern of responses rather than 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-methods or common methods bias (Cook and Campbell, 1979), which is a threat to both discriminant and convergent validity (and also to reliability; see later).

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 randomly separate construct-related items. 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

[15] Lack of random ordering of items may also explain some of the extremely high Cronbach alphas in the upper .90 range found throughout the TAM research stream.

instrument. Again, the method itself may be a major factor in how participants respond rather than a careful, thoughtful true response that reveals the true score.[16]

A method of assessing common methods bias is through 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 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 each other (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. Second order CFA is extremely rare in MIS research.

Interestingly, 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 have avoided this threat, particularly multiple methods of

[16] Methodologists have proposed quantitative techniques for analyzing single methods bias such as posed in this case. Avolio, Yammarino, and Bass (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). See also Bagozzi and Phillips (1982) and Bagozzi (1980) for use of 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.

gathering IVs and DVs, but the fact remains that few articles using one method test for the presence of this bias.

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

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

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

In the final analysis, the best heuristic for dealing with common methods bias is to completely avoid it 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. There is no real possibility that methods have any effect on the scores in this scenario.

2.3.7 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. There are many established techniques for asserting valid constructs, and many more that are evolving. What would seem to be best practice in the present time is to utilize 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, it can be tested through SEM. Testing for common methods bias is a **recommended** technique, therefore. Nomological validity is likewise a **recommended** technique, to be thought of as a supplement to conventional construct validity approaches.

2.4 Predictive Validity

Also known as "practical," "criterion-related," "postdiction," or "concurrent validity,"[17] 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

[17] 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.

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 there is a recognized theoretical connection, then predictive validity serves to reinforce the theory base (Szajna, 1994).

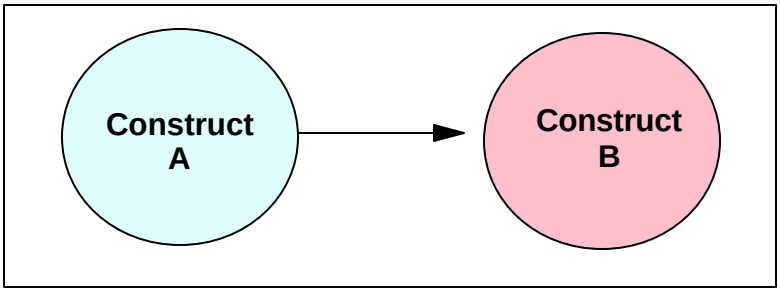


Figure 5. Pictorial Model of Predictive Validity

The widespread use of GMAT scores to predict performance in graduate studies is a case in point in the academic setting. Decision-makers have an implicit belief that constructs having to do with 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 is demonstrating 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.

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. Not discussed in Straub (1989), predictive validity could be put to better use in IS research,[18] 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's (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, which were measured at one point in time, were used to predict actual choice of a database management system, which was 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.

2.4.1 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.

2.5 Reliability

While construct validity is an issue of measurement *between* constructs, reliability is an

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

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 extent to which the measures 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 "reflect" it in correlated measures.

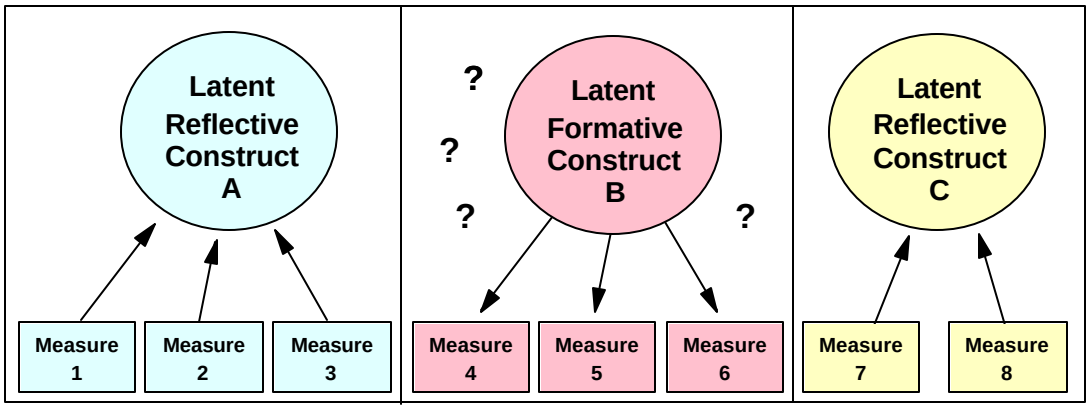


Figure 6. Pictorial Model of Reliability

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).

There are five generally recognized techniques used to assess reliability: (1) internal consistency, (2) split halves, (3) test-retest, (4) alternative or equivalent forms, and (5) inter-

rater reliability, where techniques 2-4 are considered to be more "traditional" methods. Reliability coefficients analogous to internal consistency are available also in SEM. In addition, covariance-based SEM, such as LISREL, can also be used to assess unidimensionality.

Boudreau et al., (2001) assessed the extent to which reliability was considered by researchers when developing their instruments. They discovered that the majority of the sampled publish research, that is 63%, assessed reliability. Most of this work (79%) estimated reliability through the standard coefficient of internal consistency, i.e., Cronbach's ?. Only in rare cases were other methods been used to verify 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 is regrettable as the use of additional methods to calculate Cronbach's ?, or a combination of methods, would strengthen this component of instrument validation. In the following sections, the aforementioned six types of reliability are discussed.

2.5.1 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.

Cronbach's ? 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 ? 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 should plug Spearman correlations into the K20

formulation[19] or, more closely related to the likely values of a Cronbach's ?, utilize 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 values that are very high, i.e., in the .95 and above category, are more suspect than those in the middle alpha ranges.[20] 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 has become an artifact in the measurement. As noted above, 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. 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 instance, many researchers have used Davis' TAM instrument without randomly varying the arrangement of the items composing each scale (see Table 2, above, and discussion of discriminant validity). Non-randomized ordering of questionnaire items could

[19] The formula for coefficient ? is: (k / k-1) (1- ((? $?_i^2$) / ($?_t^2$), where k = number of parts / items in the scale, $?_i^2$ = the variance of item i, and $?_t^2$ = the total variance of the scale.

easily explain the high Cronbach's ? s that are reported as well as other methods artifacts related to discriminant and convergent validity (Straub et al., 1995).

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 ? s 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).[21]

2.5.2 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 fact, there are specific cases where these techniques continue to be useful in IS research, and IS researchers need to understand why these cases justify the use of these techniques.

2.5.2.1 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

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

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). There needs to be a special purpose for using this technique, as in the case of Segars (1997), discussed later. In general, its use has been 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 ? s ranging from .62 to .90.

2.5.2.2 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,

---

[21] Nunnally reaffirmed these guidelines in his subsequent work (Nunnally, 1978; Nunnally and Bernstein, 1994)

albeit slightly lower than Davis' (1989). What is clear in the case of this validation research is that reliability of the TAM instrument had been 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. First, there is a 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.

2.5.2.3 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 that has been 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, and there is no way of determining *a posteriori* which of these tests represents the better comparative test. 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 this 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 have 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, it may be that this form of reliability may not be a first choice for IS researchers.

2.5.3 Inter-Rater Reliability

There are many occasions in empirical research when data being collected does not manifest itself in a natural quantitative form. A great deal of unstructured and semi-structured discourse in interview transcript 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, there are several unresolved issues. 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). As suggested by Miles and Huberman (1994), 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 when there are only two raters (Jones et al., 1983).[22] Miles and Huberman's

[22] 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

(1994), Landis and Koch's (1977), and Bowers and Courtright's (1984) recommendations for minimum inter-rater reliability are .70.

Vignette #8: Examples of Inter-Rater Reliability

As an example of inter-rater reliability, Lim et al.'s (1997) study of computer system learning had two independent coders score tests based on explicit instructions. These instructions described how to determine if a good explanation had been 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 provided by 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 usage or non-usage of research validities. They determined that the Kappa for their raters' coding exceeded the benchmark .70 threshold. Percentages of agreement were in the 74-100% range.

2.5.4 Unidimensional Reliability

This is perhaps one of the most important statistical tests discussed in this paper on research methods, but, alas is perhaps the least understood and certainly the least applied.

---

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.

Unidimensioanlity is a propert,y of a measurement item that states and examines that the item measures, that is reflects only one latent construct. Unidimensionality is assumed *a-priori* in many measurements of reliability, including Cronbach's ? . Gefen (2003) discusses this relationship extensively.

Techniques in covariance-based SEM can also help to determine the unidimensionality as well as 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). Contrariwise, 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 of art 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 has been long recognized as a basic assumption upon which other measures of reliability rely.[23] Unfortunately, measuring unidimensionality was extremely laborious from a statistical standpoint prior to the advent of SEM and Item Response Theory (Hambleton et al., 1984). 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 (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.

[23] It is unclear at this point as to whether unidimensionality is only a characteristic of reliability, as in Segars (1997), or whether it can also be or should be thought of as a form of construct validity. What is

Last, readers will be advised to check with Gefen (2003) in that he presents an example and step-by-step walkthrough on the use of unidimensionality in LISREL and the threats it addresses, including real data that can be used to replicate the arguments on the necessity of performing this kind of analysis. The article 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.

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.[24]

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.

Finally, 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).

---

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.

[24] The resulting two factor model with 8 items shows high fit values and also passes tests for both convergent and discriminant validity.

2.5.5 Heuristics for Reliability

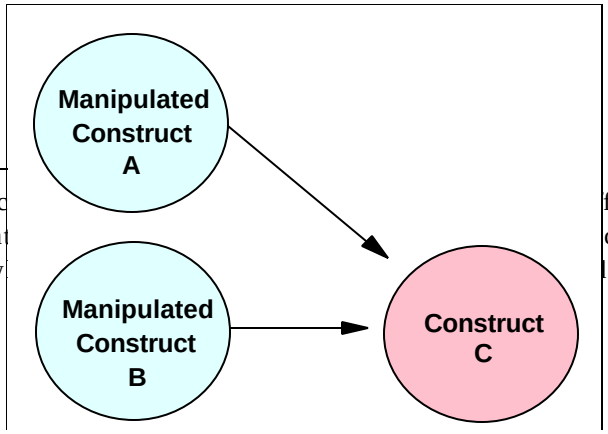
Reliability assures us that measures that should be related to each other within the same construct are, indeed, related to each other. Without reliable measures, it is difficult to see how the data can be trusted, any more than instruments lacking construct validity can be trusted. One form or another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

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

Unidimensional validity is an important new approach in the IS researcher's toolkit. Since it has not been used extensively, the field has little experience with it. We suggest that it be classified as **optional** until we gain more experience with it and understand better its capabilities.

2.6 Manipulation Validity

Manipulation validity (a.k.a. manipulation checks) is traditionally inserted into experimental procedures / tests to measure the extent to which treatments (IVs) have been perceived by the subjects (Bagozzi, 1977). As depicted in Figure 7, manipulated constructs A and B, or treatments, are the 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, there is little doubt of this,[25] but in cases where researchers are manipulating the subjects' perceptions, through experimental tasks or exercises, there can be considerable doubt.



[25] Nevertheless, physiologic fect, whereby subjects who did not get a treatment (cont cebo so the control group can be compared to those w l therefore, no effect.

Figure 7. Pictorial Model of Manipulation Checks

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 manipulation to have 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.[26] Manipulation validity is designed to ensure that subjects have, indeed, been manipulated as intended; this, therefore, is a validity that can be empirically examined.

Manipulation validity can be simple and straight-forward or they can be more complex. One common form of check is a simple question or questionnaire item on the experimental test that asks the subjects directly if they have 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 revealed that techniques such as t-test, chi-square, and ANOVA were deployed about twice as often as descriptive statistics such as counts, means,

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

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 more sophisticated technique, Gefen (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 revealed 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).

2.6.1 Heuristics for Manipulation Validity

Manipulation validity is **mandatory** for experimentation. Without these checks, the experimenters cannot be certain which subjects have been exposed to the treatments and which have not. When a subject is being treated with a physical substance, such as a drug, there is no need for manipulation validity. 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 is a recommended heuristic. It should be noted, however, that because unmanipulated subjects' responses presumably add unexplained variance, inclusion of these subjects in the dataset is a more robust testing of the hypotheses and some authors may choose to retain them for this reason.[27]

2.7 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 newly developed techniques that assist in establishing statistical conclusion validity in the last decade. These techniques take different approaches to establishing whether, statistically, there is a "causal 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 as to call for special treatment in this paper. There are two types of SEM: 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 are not explicitly stated in

[27] The counter argument to pooling the manipulated and unmanipulated subjects 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.

the model (see Gefen, Straub, and Boudreau, 2000, for a detailed discussion and comparison.) Their effective use in IS research is the underlying issue. These SEM techniques are now widely used in the top IS journals (Gefen et al., 2000).

Vignette #11: Examples of Statistical Conclusion Validity

There are any number of examples of IS authors employing statistical conclusion validity, specifically justifying the specific type of tool used based on its inherent distribution assumptions and susceptibility to 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 in length why it is more appropriate than LISREL for their specific data and their predictive rather than confirmatory approach. As to Taylor and Todd (1995), they explained why they preferred LISREL for their analysis because of its ability to compare alternative models dealing with well established theories.

2.7.1 Heuristics of Statistical Conclusion Validity

Statistical conclusion validity is mentioned briefly here for the sake of completeness. This technique has received the single most attention in management research (Scandura and Williams, 2000), and 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.

3. Guidelines For Research Practice

Although it has 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 are essentially pragmatic measures indicating patterns of behavior that appear to be acceptable within the IS scientific community. There is no recognized means of verifying 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 that the results being reported could be false, then a new alpha level would become the rule of thumb. There is no mathematical or other means for establishing these levels (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 tradition 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. Please note that 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 utilize 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 below in Tables 3a through 3d.

Validity	Techniqu	Heuristic	Source
CONSTRUCT VALIDITY (MANDATORY) ✍✍ *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 ?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 ?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 (1998c) 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 3a. Mandatory Validities

Validity Component	Technique	Heuristic	Source
RELIABILITY (MANDATORY) ✍✍ *Internal consistency* **(Recommended over other tests, where appropriate)**	Cronbach's ? ; correlations; SEM reliability coefficients	Cronbach's ? 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* **(Where appropriate)**	Coefficient kappa; correlations; percentages;	Coefficient Kappa > .70	Landis and Koch (1977) Miles and Huberman (1994)
MANIPULATION VALIDITY **(Where appropriate)**	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	Perdue and Summers (1986)

Table 3b. Mandatory Validities (Continued)

Validity Component	Technique	Heuristic	Source
CONTENT VALIDITY ✍✍ **(HIGHLY RECOMMENDED)**	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 3c. Highly Recommended Validities

Validity Component	Technique	Heuristic	Source
PREDICTIVE VALIDITY (OPTIONAL)	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 ? >.60 / .70 and < .95	
✍✍ Test-retest	Same as internal consistency	Cronbach's ? >.60 / .70 and < .95	
✍✍ Alternative forms	Same as internal consistency	Cronbach's ? >.60 / .70 and < .95	
✍✍ Unidimensional reliability	Model comparisons	Model comparisons favor unidimensionality	Segars (1997) Gefen et al. (2000) Gefen (2003)

Table 3d. Optional, but Recommended Validities

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 the present paper. The Straub 1989 guidelines will, therefore, be subsumed into "Guidelines for the Year 2003 and Beyond," immediately below.

3.1 Guidelines for the Year 2003 and Beyond

Two broadly stated guidelines emerge from the present study:

1. *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 following rigorous standards for validation. To pointedly reiterate and elaborate on these:

   ✍✍ ***Content validity*** is **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*** is **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*** is **optional**: Establishing *predictive validity* is useful for mature research streams.

✍✍ ***Reliability*** is **mandatory**: Establishing *reliability* is a necessary practice, with Cronbach's ? 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*** remains **optional** in spite of its growing importance. At the moment, there is little understanding of the techniques in IS research. Over time it could become **mandatory** because all reliability measures, including Cronbach's ? , 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 is **mandatory**: Establishing *manipulation validity* is a necessary practice for determining the validity of treatments (independent variables) in experimentation. With regard to other aspects of instrumentation, experiments should be subject to the same validity standards as other research methods.

✍✍ ***Statistical conclusion validity*** is **mandatory**: Establishing *statistical conclusion validity* is essential for all quantitative, positivist research.

2. *Innovation in Instrumentation*

✍✍ ***Use of previously validated instruments*** is **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*** is **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, 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 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.

3.1.1 CRITICISM OF THESE VALIDATION GUIDELINES WELCOME

Before discussing cases in which there might be contingencies for our validation guidelines, we wish to stress that all such validation guidelines are owned by communities-of-practice and should not be the provenance of methodological "experts" or others in positions of power. 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.

3.1.2 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 differentially applied. 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 4), 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 have pursued in this article is that the kind

of highly theoretical, confirmatory work that most frequently appears in the top ranked IS journals will require the rigor of full validation and 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 cogently argue 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 4, 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 of these 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 and so predictive validity or nomological validity are beyond what many readers might require.

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	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
3	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

4	Non-theoretical work	Examining the nature of a phenomenon through descriptive statistics primarily, or with highly exploratory hypothesis testing	Content validity[28]; predictive validity
5	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

Table 4. Contingencies for Where Validities Apply

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 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.

4. 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 exemplar of how new instruments can be developed is offered next. Smith et al. (1996) validated their information privacy instrument through a

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

judicious choice of most of the validation techniques we have discussed. Each of these techniques will be briefly described below, to show the extent of the validation undertaken.

Content Validity. First, five different groups were asked to assess the content of dimensions that the authors proposed for information privacy. Initially, three experts in the privacy area were given 72 preliminary questionnaire items. 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. Thirty-two remaining items were next evaluated by 15 corporate employees, followed by 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 utilized.

Predictive Validity. The practical test of the instrument was to determine whether its measures would correlate highly with criterion 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.

5. CONCLUSION

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 (see

http://www.endnote.auckland.ac.nz/ for a bibliography developed entirely to methodology; this represents just a starting point for more work in this area). It is clear that the field has a wide variety of methodology specialists who are capable of articulating the principles that guide their practice. To encourage this dialogue could, one might argue, "should" be a worthy goal of IS journals. The quality of our science should be *sine qua non*, "without which nothing."

Other fields have taken such an introspective look at their research strategies and extent of validation (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 positivist approach that one takes so that other 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 have been proffered. It is critical that the gatekeepers of the field, as represented by the journals and conferences, raise the level of awareness of the entire community by insisting on the standards offered here or others convincingly presented.

6. REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gefen, D. "Building Users' Trust in Freeware Providers and the Effects of this Trust on Users' Perceptions of Usefulness, Ease of Use and Intended Use," in: *Computer Information Systems*, Georgia State University, Atlanta, GA USA, 1997.

Gefen, D. "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) 2000, pp 65-79.

Gefen, D. "Assessing Unidimensionality Through LISREL: An Explanation and an Example," *Communications of AIS* (12:2), 23-47 2003.

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

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

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

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

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

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

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

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

Jenkins, A.M. "Research Methodologies and MIS Research," in: *Research Methods in Information Systems,* Enid Mumford et al. (ed.), Elsevier Science Publishers, Amsterdam, Holland, 1985, pp. 103-117.

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

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

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

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

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

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

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

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

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

MacCallum, R.C., Roznowski, M., and Necowitz, L.B. "Model modifications in covariance structure analysis: The problem of capitalization on chance," *Psychological Bulletin* (111:3) 1992, pp 490-504.

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

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

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

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

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

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

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

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

Nunnally, J.C., and Bernstein, I.H. *Psychometric Theory, Third Edition* McGraw-Hill, New York, 1994.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Simon, S.J., Grover, V., Teng, J.T.C., and Whitcomb, K. "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) 1996, pp 466-490.

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

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

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

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

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

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

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

Thompson, R., Barclay, D.W., and Higgins, C.A. "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)) 1995, pp 284-324.

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

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

Woszczynski, Amy B. and Michael E. Whitman, "The Problem of Common Method Variance in IS Research," In *The Handbook of Information Systems Research*, M. E. Whitman and A. B. Woszczynski (Ed.), Idea Group Publishing, Hershey, PA USA, 2004, 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$(? ?_i^2)/((? ?_i^2) + (? (1-?_i^2))$, 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 ? :** 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:** 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:** 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 chi-square 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.