

How to Tie a Construct to Indicators: Guidelines for Valid Measurement

Completed Research Paper

Sander Paul Zwanenburg
The University of Hong Kong
Pokfulam Road, Hong Kong
info@sanderpaul.com

Abstract

Invalid measurement of constructs in information systems research often remains undetected and can lead to false conclusions. The prescriptive literature on measurement has led to a better understanding of the sources of error in various areas, including conceptual modeling, common method bias, and estimation procedures. It has also called for heterogeneity in indicators to overcome sources of error associated with each indicator specifically. It has not led, however, to widespread measurement practice that takes these separate insights into account. This paper aims to facilitate this by integrating insights from the literature. It complements extant guidelines on the development of measurement with a typology of the ways to tie a construct to its indicators. It demonstrates the recommendations with an empirical illustration. This, I hope, will lead researchers adopt more heterogeneous indicators, allowing them to measure their constructs with better confidence in validity.

Keywords: measurement, indicator, construct, heterogeneity, guidelines, development, validity

Introduction

The validity of measurement has been a cause of much concern in the field of Information Systems. Various reviews have found that many researchers do not sufficiently demonstrate the validity of their measurement (Boudreau et al. 2001), especially when using formative indicators (Ringle et al. 2012). Although a lack of validation does not imply a lack of validity, studies in the field suggest a need for caution when relying on measurement that lacks proper validation. Studies on method bias, for example, found that researchers have dealt with this bias insufficiently (King et al. 2007; Woszczyński and Whitman 2004), even though it can explain much variance in measurement estimates (Sharma et al. 2009). They have also shown that statistical techniques used to control for method bias can be ineffective (Chin et al. 2012). Misspecification of measurement models is another threat to the validity of measurement (Aguirre-Urreta and Marakas 2012; Jarvis et al. 2012; Petter et al. 2012). Petter et al. (2007) found that 30% of the measured constructs in articles published in 2003-2005 in two leading journals of the field were modeled as reflective but should have been formative. An ongoing debate on formative measurement suggests that researchers may continue misspecify their measurement models (e.g. Aguirre-Urreta and Marakas 2013; Aguirre-Urreta and Marakas 2014; Kim et al. 2010; Rigdon et al. 2014). Taken together, these studies show that the measurement practices in field of IS still have much room for improvement, both in establishing and demonstrating validity.

This is not for lack of attention in the literature. Information Systems journals have provided guidance on minimizing method bias (Burton-Jones 2009), specifying formatively measured constructs (Petter et al. 2007), interpreting measurement results (Cenfetelli and Bassellier 2009), assessing hierarchical construct

models (Wetzels et al. 2009), and validating measurement (Boudreau et al. 2001; MacKenzie et al. 2011). They have also set examples, for example in measurement over time (e.g. Venkatesh et al. 2003), with multiple dimensions (e.g. Segars and Grover 1998), and through multiple methods (e.g. Ortiz de Guinea and Webster 2013). IS researchers can also find a multitude of recommendations outside of their field, in books and in the journals of reference disciplines including psychology, management, and marketing.

How is it possible that so much guidance is available while there is still much room for improvement? I concur with MacKenzie et al. (2011): “a [...] likely possibility is that there is simply so much work on the topic of scale development and evaluation that it is difficult for researchers to prioritize what needs to be done” (p. 294). This idea has led MacKenzie et al. (2011) to integrate guidance on an array of topics of measurement into a set of recommendations for the development of construct measurement. They have taken an excellent step forward from the guidance provided in Churchill’s (1979) seminal article.

With this paper I continue this line of work. This paper integrates a rationale for heterogeneity in measurement with guidance on the development of measurement. While the literature has provided insights into various threats to validity and called for more heterogeneity in measurement, guidance on the development of measurement has typically assumed a construct is to be measured through a single means. It commonly assumes the use of a single measurement model, a one-off questionnaire with a single rater, and a single method for estimation. It also assumes indicators are either reflective or formative. In this paper, we propose a typology of all the ways to tie a construct to its indicators. This, I hope, would help researchers identify opportunities to measure their construct with heterogeneous indicators.

The Development of Measurement

The goal of measurement is to infer the position of entities on a construct (Markus and Borsboom 2013). As shown in panel A of Figure 1, this process of measurement consists of (1) conceptually tying the construct to indicators, (2) operating these indicators to produce records, (3) mathematically combining these records to calculate estimates, and (4) attributing these estimates to the construct.

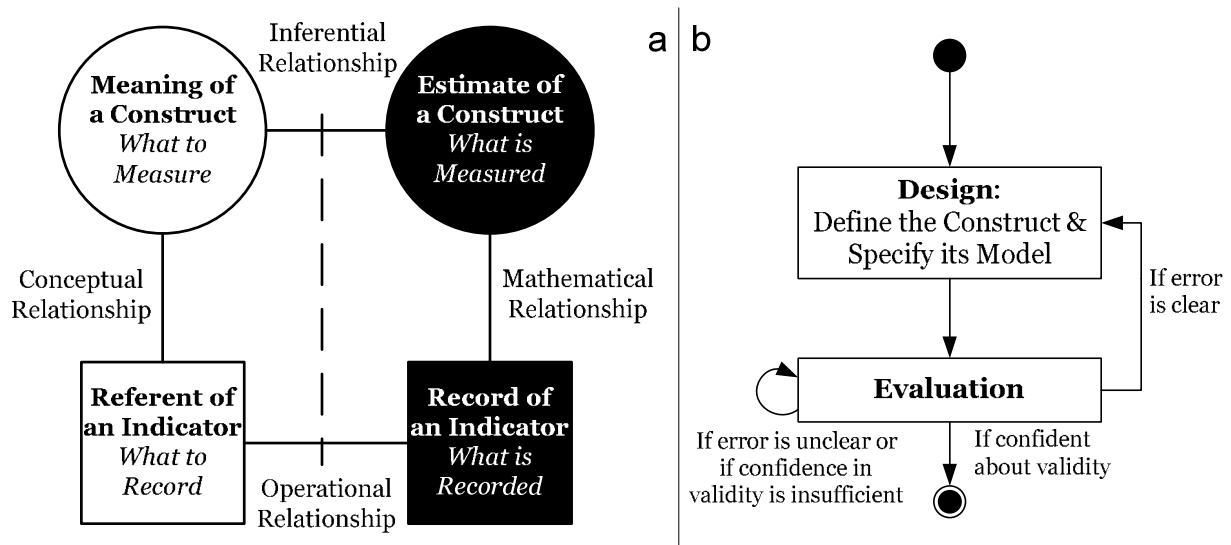


Figure 1. Models of the Structure (panel a) and the Development of Measurement (panel b)

This last measurement inference is valid when these estimates, i.e. what is measured, matches the meaning of the construct, i.e. what is to be measured (Markus and Borsboom 2013). This depends, critically, on the conceptual, operational, and mathematical steps that preceded it. Conceptually, a construct may not be equivalent to the combination of items because, for example, indicators correspond to only some parts of a construct, or to effects of a construct that are also effects of something else (e.g. Petter et al. 2007). Operationally, what is recorded can deviate from what was to be recorded because, for example, a respondent misinterprets a question or lies about it (e.g. Podsakoff et al. 2003). Mathematically, combining records can undermine validity when this is inconsistent with an understanding of the conceptual and

operational steps that have led to the records. For example, the local independence assumption of confirmatory factor analysis may not be consistent with the conceptual relations of the indicators (e.g. Meredith 1993). Understanding these sources of error is thus key to develop valid measurement.

The development starts with an initial understanding of the meaning of a construct; its desired end state is the measurement of the construct with confidence in its validity. The transformation from start to end can be complex and typically requires an iterative cycle of *design* and *evaluation*, as depicted in panel B of Figure 1. ‘Designing’ here refers to both defining the construct and specifying its measurement model. This combination may be evaluated through simple thought experiments or through elaborate programs of data collection and validity tests. The outcome may lead researchers to revise the measurement model or even to redefine the construct (Churchill 1979; MacKenzie et al. 2011). This iterative cycle continues until sufficient confidence is obtained that the construct is measured with a satisfactory level of validity.

Defining the Construct

While it is obvious that the definition of a construct is of critical importance to the design and evaluation of measurement, it often receives insufficient attention (MacKenzie et al. 2011). Defining a construct requires knowing what to measure and how to describe that, which can be confusing (MacKenzie 2003). A construct’s name, i.e. its label or term, can also denote other constructs as it can carry multiple meanings. One meaning can be described in different ways; a construct can be defined in different languages and syntaxes. Definitions of a construct can also highlight different aspects of its meaning. As long as they are consistent, multiple definitions of the same construct help specify its meaning, i.e. they demarcate it (Barki 2008; Goertz 2006; MacKenzie et al. 2011). These definitions must be consistent; when two definitions are inconsistent they define different constructs.

Recognizing good definitions can be difficult. The meaning of a construct can never be fully specified, in the sense that any description of a construct is to some degree ambiguous (Kaplan 1964; Van de Ven 2007). Sometimes ambiguity can obstruct the validity of inferences made about the construct. Hence, while ambiguities are inevitable, it is imperative to remove those that can hinder the goals of the research inquiry.

Both defining and measuring a construct relies on the relationships between a construct and other concepts, such as its causes, effects, constituent parts, and dimensions (MacKenzie et al. 2011). While defining, these relationships help set or fix the target conceptually, while in measuring they help aim and hit the target operationally.

Specifying a Measurement Model

To specify a measurement model means to translate a web of understanding about a construct into a plan of a linear string of operations to measure it. It consists of generating indicators and specifying how they relate to the construct. Indicators refer to what is to be recorded in order to measure the construct.

Constructs vary widely in how many indicators they connect to. While sometimes a construct is measured with a single indicator, such as most measurements of gender and age, measurement often involves multiple ones (Bergkvist and Rossiter 2007). They can indicate their construct either directly or indirectly through intermediate, latent indicators (Edwards 2001; Law et al. 1998; Polites et al. 2012). These latent indicators are also called ‘sub-constructs’ because they have the same measurement properties as constructs: they too can be measured directly or indirectly. Non-latent indicators are also known as ‘manifest indicators’ or items. When they are operated, they yield one record for each instantiation of the construct (such as for each individual, or firm-year). These records can then be combined to produce estimates of the sub-constructs and ultimately the construct itself.

Layers of indicators thus form a tree (or ‘hierarchy’) of conceptual relationships. The left half of the diagram in Figure 2 (with the open nodes) depicts such a tree while the right half (with the nodes filled) depicts the associated records and estimates. The figure also provides some ostensive definitions of some key terms. For simplicity, in this paper I use ‘construct’ to refer to any construct, at any level; I use ‘indicator’ in the context of the construct it *directly* indicates.

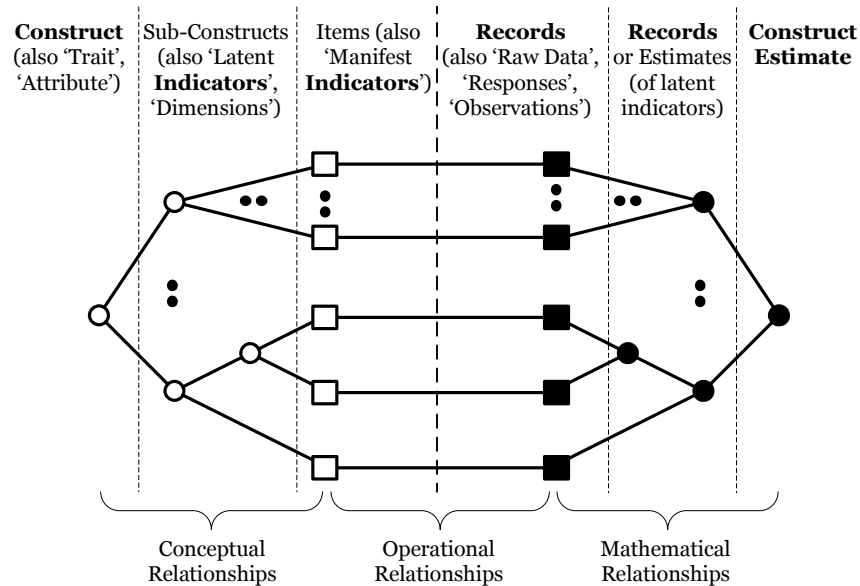


Figure 2. A Network Diagram of Measurement and its Terminology

The hierarchical structure highlights the idea that the connection between a construct with its items can be decomposed into one or more logical steps. In each, a construct (or sub-construct) is translated into a combination of (direct) indicators.

This can be done in a plethora of ways. The literature has predominantly focused on two broad categories of these ways: generating *reflective* indicators and generating *formative* ones, as shown in Table 1. Much literature exists that elucidates how and when these two types can ease measurement (Blalock 1964; Bollen and Lennox 1991; Bollen and Bauldry 2011; Diamantopoulos 2011; Diamantopoulos et al. 2008; Diamantopoulos and Temme 2013; Diamantopoulos and Winklhofer 2001; Edwards 2001; Edwards 2011; Edwards and Bagozzi 2000; Howell et al. 2013; Kim et al. 2010; Lee et al. 2013; MacKenzie et al. 2011; Markus and Borsboom 2013; Petter et al. 2007; Rigdon 2013).

Aspect	Type of Indicators	
	Reflective	Formative
Their target	The set of indicators should cover or represent all content areas, dimensions, or facets of the construct (Churchill 1979; Clark and Watson 1995; Loevinger 1957; MacKenzie et al. 2011; Netemeyer et al. 2003)	The set of formative indicators should cover all content areas, dimensions, or facets of the construct (Bollen and Lennox 1991; Diamantopoulos and Winklhofer 2001; Kim et al. 2010; MacKenzie et al. 2011; Netemeyer et al. 2003). “Breadth of definition is extremely important to causal indicators” (Nunnally and Bernstein 1994, p484).
	Capture a domain that is slightly broader than that of the ‘expected’ construct (Clark and Watson 1995; Loevinger 1957)	
	“The items should not venture beyond the bounds of the defining construct” (DeVellis 2003, p64)	
	The items should be distributed according to the relative importance of the facets of the construct (Haynes et al. 1995)	
Metaphor of their generation	Select items as if drawing a <i>random</i> sample form a hypothetical universe of items (Cronbach 1951; DeVellis 2003; Diamantopoulos and Siguaw 2006; Nunnally and Bernstein 1994). “The items must represent a reasonable sample of items tapping the domain of the construct” (Netemeyer et al.	Select items as if conducting a census (Bollen and Lennox 1991; Diamantopoulos and Siguaw 2006; Netemeyer et al. 2003).

	2003, p93) “In generating an item pool, an important goal is to <i>systematically</i> sample all content areas of the construct.” (Netemeyer et al. 2003, p95); a construct is to be subdivided based on theoretical and empirical utility (Spector 1992).	
Differences and similarities across indicators	They should share a common theme (Kim et al. 2010). They should express the same idea in different ways, e.g. by using different words and grammatical structures (DeVellis 2003; Nunnally and Bernstein 1994). “The researcher probably would want to include items with slightly different shades of meaning” (p68 Churchill 1979)	Each item should cover one aspect; “omitting an indicator is omitting a part of the construct” (p308 Bollen and Lennox 1991). They should not share the common theme (Kim et al. 2010).
	“An ideal item in a test that measures a broad trait is one that has a relatively high correlation with the sum of all items in the test (minus itself) and a relatively low average correlation with the other items” (p366 Epstein 1983).	To avoid multi-collinearity, they should not be too highly correlated (Diamantopoulos and Winklhofer 2001).
Sources of inspiration	Useful sources include literature reviews, theoretical deductions, experts, representatives of the relevant population, and extant measures (Churchill 1979; Haynes et al. 1995; Kim et al. 2010; MacKenzie et al. 2011)	Useful sources include literature reviews, theoretical deductions, experts, representatives of the relevant population, and extant measures (Kim et al. 2010; MacKenzie et al. 2011)
Their initial number	“A larger number is preferred, as overinclusiveness is more desirable than underinclusiveness” (Netemeyer et al. 2003, p102). “It would not be unusual to begin with a pool of items that is three or four times as large as the final scale [...] The larger the item pool, the better” (DeVellis 2003, p66)	

Table 1. Extant Guidelines for Generating Indicators

The attention the two types of indicators have received may have overshadowed the full breadth of how a construct can in principle connect to its indicators. For example, most guidelines do not consider multiplicative indicators (Law et al. 1998; MacKenzie et al. 2011), subtractive indicators (Klein et al. 2009), indicators that correct other indicators (Podsakoff et al. 2003), indicators that correspond to the same question over time (Csikszentmihalyi and Larson 1987), or the same question asked to different informants (Burton-Jones 2009). All such indicators may prove useful for the development of measurement as they may be differentially sensitive to error. An integral typology of ways to tie a construct to indicators may therefore help researchers in their efforts to validly measure their construct of interest. In the next subsections, we present such a typology by introducing its two dimensions: the conceptual distinction between indicators, and their composition in relation to the construct.

The Distinction between Indicators

The logical basis of translating a construct into a combination of direct indicators is either substantive or methodical, as illustrated in the columns of Table 2. It is *substantive* when the indicators refer to meaning (or ‘content’) that stands in a relation to the meaning of the construct. For example, they may refer to constituent parts of the construct or its effects. The basis for specifying substantive indicators is a conceptual understanding of the construct: how does the construct behave in time and space? This understanding may include the construct’s dynamics (what causes it, what is it caused by) and its structure (what its parts are, what it is part of, what its dimensions are, and what it is a dimension of). Each of these concepts, namely causes, effects, parts, the whole the construct is part of, dimensions, and the whole the construct is a dimension of, are potential substantive indicators of the construct (Goertz 2006; MacKenzie 2003; MacKenzie et al. 2011). These substantive indicators can be modeled as reflective when they refer to local-

ly independent effects of a construct (i.e. when a construct causes effect i independently of it causing effect j, for any i and j). They can be modeled as formative when they refer to parts that combine additively to make up the construct. They are neither reflective nor formative when they require multiplication or subtraction (MacKenzie et al. 2011; Polites et al. 2012).

Indicators are *methodical* when they refer to different ways of capturing the construct. For example, some indicators refer to differently worded statements in a questionnaire that aim to capture the same construct (Churchill 1979), while entire instruments of a construct can also function as separate indicators. What drives the specification of methodical indicators is an operational understanding of how the construct can be captured: what are the methods of observation and what are their respective pitfalls? What traces does the construct leave: human memories, technical recordings or both? How does an attempt at accessing a trace distort the way it is captured? Methodical indicators are often (but need not be) reflective as they indicate the same construct (but through different means) – a change in the position on a construct would be expected to be reflected in each indicator (MacKenzie et al. 2011).

Composition of Indicators	Conceptual Distinction of Indicators	
	Substance	Method
The referents of indicators overlap each other with the entire meaning of the construct. They capture the construct in their own (deficient) way. Differences in deficiencies attenuate their impact on validity.	The indicators refer to different substance, having in common an overlap with the entire meaning of a construct. They are reflective. Examples: Indicators refer to different <i>independent effects</i> (i.e. the construct causes multiple effects in conjunction), or to different <i>manifestations</i> of a construct, such as different symptoms of a disease, or behaviors of a personality type.	Each indicator captures the entire construct through different methods. An aspect of measurement distinguishes the indicators. They are reflective. Examples: indicators that correspond to differences in <i>how</i> a question is asked (Churchill 1979; Netemeyer et al. 2003), <i>to whom</i> (using multiple informants for measuring an entity; e.g. Kumar et al. 1993, Burton-Jones 2009), <i>when</i> (longitudinal measurement of stable constructs; e.g. Os et al. 2013, Csikszentmihalyi and Larson 1987), or <i>where</i> (location-based assessment); indicators correspond to different <i>instruments</i> or <i>estimation techniques</i> .
The referents of indicators complement each other. Each refers to the shortcoming of the combination of the other indicators.	Each indicator captures the substantive difference between the construct and the combination of the other indicators. When their records are to be summed, they are formative; otherwise they are neither reflective nor formative. Examples: summative <i>parts</i> that make up the whole, <i>dimensions</i> that combine multiplicatively to form the construct (e.g. length and weight form obesity, probability and impact form risk), <i>causes</i> , and <i>disjunct effects</i> (when the construct causes either Effect A or Effect B).	One indicator corrects the method error of another indicator. They are neither reflective nor formative. Examples: measuring the weight of a liquid by weighting it inside a container and by weighting the empty container; indicating a construct by self-reports and by a correction for bias in self-reports (Harman 1976; Nederhof 1985; Podsakoff et al. 2003; Podsakoff et al. 2012).

Table 2. A Typology of Indicators

The Composition of Indicators

Any combination of indicators to measure the construct falls into one of two categories, as illustrated in the rows of Table 2. First, the referents of indicators help measure the construct by *overlapping* with each other and with the entire meaning of the construct. They may differ in either substance or method. The

logic of overlapping indicators is that they will be differentially sensitive to sources of error, such that combining them should produce a good estimate of the construct. The more heterogeneous they are, the less sources of error they share in common, and the higher the validity they can allow for. For example, a construct may be measured through indicators that refer to its multiple effects. Capturing these effects may be deficient in different ways: perhaps the construct does not cause effect A under certain conditions; effect B is sometimes caused by something else entirely; and effect C does not leave many traces in human memory. Differences in the sources of error across indicators can be leveraged to attenuate the effect of these sources on the validity of the measurement estimate. Thus, the degree to which the indicators combine to valid estimates of the construct depends on the errors associated with what they have in common.

Alternatively, the referents of the indicators may *complement* each other. Each indicator may refer to the difference between the construct and the combination of the rest of the indicators (i.e. their shortcoming). A straightforward example is Carlson and Grossbart's (1988) approach to measuring the amount of TV parents watch with their children throughout the week: they cut up the week into weekdays, Saturdays, and Sundays. As an indicator, the amount of watching TV on weekdays falls short conceptually as it does not capture the weekend. The other indicators, however, refer precisely to this shortcoming. A methodical example is measuring a sensitive behavioral construct with a biased self-report indicator of this behavior and an indicator that refers to the bias the self-report suffers from (Harman 1976; Nederhof 1985; Podsakoff et al. 2003; Podsakoff et al. 2012). While this bias indicator has nothing substantively to do with the construct, logically, it is as much an indicator of the construct as the weekday TV watching indicator is of weekly TV watching. They both complement the set of other indicators by isolating its shortcoming.

The Information of Indication

Regardless of their composition and distinction, combinations of indicators may carry different degrees of information about the construct. Some indicators fully inform the value of the estimate. For example, data on weekday, Saturday, and Sunday TV watching combines in only a single mathematical way to produce an estimate of weekly TV watching. In many cases, however, the relationship between a construct and its combination of indicators is less evident. There may be no clear conceptual answers to questions such as: when does a construct cause the effects its indicators refer to; is a set of indicators that refers to a construct's parts complete; does an indicator refer to a cause, effect, or both; and how to combine the answers by different informants?

Information may also lack in the structure of the conceptual relationships. While conceptual grounds may inform how a construct can be cut up according to both its substance and its method, they may fall short in informing what sequence is best.

A lack of information in the conceptual relationships often increases the permissible mathematical combinations of the indicators. A range of parameter values, like those of weights or factor loadings, may be consistent with the definition of the construct. This conceptual leeway is commonly exploited by testing multiple mathematical implementations of the measurement model and choosing the one that satisfies an objective function. While this may be seen as a correction for inevitable error in measurement, it can come at a theoretical price, since it creates artificial variability in how constructs are mathematically operationalized across studies. Factor loadings, for example, are typically re-estimated for every study, a practice that has raised criticism (Rigdon 2013). In my view, these concerns merit further methodological inquiry before guidelines can be provided. Generally, however, measurement is best based on relationships that are well-understood and that involve little to no conceptual ambiguities.

An Initial Model

An understanding of (1) how indicators can be distinguished, (2) how they can work together to measure a construct, and (3) how much information they carry should be of general help in identifying the different ways to measure a construct. More construct-specific help for specifying an initial model may be found by consulting prior literature and theories, reviewing extant measures, asking experts, using focus groups, and conducting explorative surveys (Churchill 1979; MacKenzie et al. 2011; Netemeyer et al. 2003). Some researchers have proposed to simply start with a one-indicator model, in which the item aims to directly capture the construct (Netemeyer et al. 2003), and then evaluate it.

Evaluation

An evaluation of a measurement model, or part of it, aims to detect its flaws and establish its validity. This can be done in many ways, from simple thought experiments about a narrow part of the model to complicated programs of applying and analyzing the entire model of the construct along with that of others (MacKenzie et al. 2011). Intuitively, when an evaluation detects flaw, it can aid the design of measurement. When it does not, it raises confidence in validity. As depicted in the right panel of Figure 1, it may lead to new tests until validity of measurement is demonstrated with sufficient confidence.

Similar to evaluating other technology like machines or programming code, evaluating measurement can follow different procedural strategies. One can choose to first apply a test to the entire system, and if unsuccessful, test its sub-systems, and so on, until the detected error is clear enough to respond to. Alternatively, one can test bottom-up: evaluate specific parts first and then, if successful, work up toward the entire system. Similarly, one can vary the strength of the tests, first trying basic test to see if a system works in principle before testing how it does in realistic conditions, or the other way around.

Which approach is chosen is often guided by the confidence in the system and its parts, and the costs of running tests. Most costs of evaluating the validity of measurement typically lie in the collection of data. Therefore, if costs are a concern, one can execute cheaper and faster tests without data first such that more confidence is gained in the validity of measurement before performing costlier data-dependent tests.

Table 2 provides an overview of the validity tests that can be performed with different sources of evidence, along with references to literature with more details. Many of the data-dependent tests involve a comparison between the data and the expectations. Typically, constructs with indicators that are overlapping involve more expectations since they should all reflect the construct by definition. Whether other indicators can also be expected to correlate depends on them having common causes.

Sources of Evidence	Relationship to Test	
	Operational (for each manifest indicator)	Conceptual (for each latent construct)
Own thought experiments (testing for face and content validity)	Will operating the item yield observations that correspond to the meaning of that item? What kind of disturbances may occur? How does processing one item influence the process of another? If a question is asked, is comprehending and answering it easy enough? What will respondents feel and think when they answer it? See Tourangeau et al. (2000) for the psychology of survey responses, Dillman (2000) for questionnaire design, and Podsakoff et al. (2012) for potential method biases.	Is the specific combination of indicators consistent with the meaning of the construct? Is it possible to think of a case in which values on the indicators imply a construct estimate that does not match the meaning of the construct? Can I infer the position on the construct given plausible values on its indicators? Are there any sources of error that influence all or a majority of a construct's indicators?
Judgments by others (testing for face and content validity)	What do experts think about applying the items? What do participants think about them? What drives their response to the item? See Dillman (2000) for advise on conducting participant interviews in pretests.	What do others think about the match between the indicator and the construct? Do they see an indicator in a relation to the meaning of the construct in the way it is modeled? Do they see the indicators represent the entire construct, or do they see relevant aspects being ignored? See MacKenzie et al. (2011) for advise on conducting a rigorous content validity test.

Sample data from the model (testing for internal validity)	Do observations match expectations? For each item, does the mean, the variance, and the distribution of the observations make sense? Do response patterns to pairs of items match expectations? Do the inter-item correlations match expectations? Is there an item that stands out in (lack of) correlation with the items it should correlate with?	Do estimates match expectations? Do their means, variances, and distributions make sense? For those indicators that should capture the same construct, do they co-vary as expected (internal consistency; average variance extracted, coefficient alpha, composite reliability, goodness of fit)? When multiple measurements are taken of a stable construct, is test-retest reliability appropriate? See for further guidance Fornell and Larcker (1981), MacKenzie et al. (2011), Bollen and Lennox (1991), and Bollen (1989).
Sample data from the model in relation to data of other variables or samples (testing for external validity)	Do differences of observations across samples match what could be expected from the differences of the samples? Do response patterns of an item correspond to those of external variables in expected ways?	Do estimates correlate with other variables as expected? How well do combined models fit the data? See MacKenzie et al. (2011) for guidelines on conducting test of known-group differences, experimental manipulation, discriminant and convergent validity, nomological validity tests, and multiple samples.

Table 3. Evaluating Operational and Conceptual Relationships of Measurement

It is often difficult to draw conclusions based on quantitative test results. If a construct explains 50% of variance in the indicators, and the indicators’ average correlation with external variables is .44, can we conclude that validity or lack thereof is demonstrated? While many cut-off values are reported in the literature, ultimately, whether these values indicate validity or not depends on the many aspects of the relationships between a construct and its indicators. They include the common methodical aspects across indicators. Table 4 provides examples of these implications for the interpretation of different values of internal consistency metrics (e.g. Cronbach’s alpha, Average Variance Extracted, Composite Reliability) for homogeneous and heterogeneous indicators. Lower cut-off values of internal consistency metrics may thus be appropriate for more heterogeneous indicators. Generally, conclusions of validity should be based on a broad range of evidence in light of the specific construct of interest.

Level of Internal Consistency (IC)	Multiple Reflective Indicators	
	Homogeneous	Heterogeneous
High	High IC is ambiguous: it may mask common method bias, and other sources of error common to the indicators.	High IC is clear: it is evidence that the records of the indicators share in common the substance of the construct.
Low	Low IC is clear: it is evidence that the records do not overlap much with the construct.	Low IC is ambiguous: it may mask validity, as the influence of sources of error specific to indicators may be alleviated through combination.

Table 4. Interpreting Values of Internal Consistency with Different Indicators

Redesign

An evaluation may lead researchers to change their measurement model and even to redefine their construct. Attempts to measure a construct often reveal that its definition is ambiguous. Depending on the broader context of inquiry, resolving these ambiguities can be critical (MacKenzie 2003; Van de Ven 2007). This may be done by adding details or using more specific words while ensuring that changes do not cause violations to the theoretical or conceptual underpinnings of the construct.

Depending on the nature of the error(s), revisions to the measurement model can take various forms, as illustrated in Table 5. They involve the elimination of an indicator (i.e. a ‘purification’ of the model), a manipulation of an indicator (i.e. an ‘adjustment’ of the model), or the insertion of a new indicator. As the three right-most diagrams in the table show, new indicators can be inserted at one or more levels, resulting in ‘refinements’, ‘extensions’, or ‘expansions’ of the model.

The first four types of change in the table are relatively straightforward to implement or have received considerable attention in the literature (Clark and Watson 1995; Dillman 2000; Haynes et al. 1995; MacKenzie et al. 2011; Tourangeau et al. 2000). In this paper, I will provide special guidance on responding to the right-most scenario in the table. It describes a prevalent problem, as measurement often relies on a single view of a construct, such as a cause of its effects or the sum of its parts, and a single approach to capturing it, often through multiple questions in a one-off questionnaire administered with a single rater. This renders its measurement validity sensitive to the sources of errors that are specific to that view and approach (Nunnally and Bernstein 1994; Spector 2006).

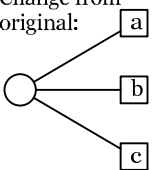
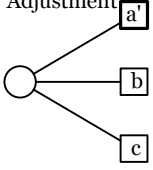
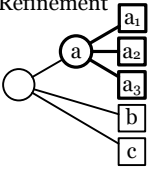
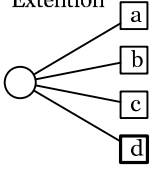
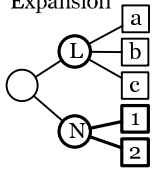
Specificity of Deficiency:	A Single Indicator			A Set of Indicators	
	Indicator is deficient but not critical to the measurement of the construct (see MacKenzie et al. 2011).	Indicator is ambiguous and can be repaired, e.g. by clarifying a question or its introduction (see e.g. Dillman, 2000)	Indicator is too complicated to relate to one observation. E.g. a question is double-barreled or requires too much evaluation.	A set of indicators is incomplete or too insensitive to the construct, e.g. by not tapping into all aspects of the construct.	All indicators suffer from the same deficiency, because of the underlying view of the construct or its method of measurement.
Change from original:					
	Purification	Adjustment	Refinement	Extention	Expansion

Table 5. Revising a Measurement Model

Expansion

When sources of error threaten the validity of all indicators, expanding the measurement model with another set of indicators can help reduce the sensitivity to those errors. When a set primarily suffers from a specific source of error that can be isolated with another set of indicators, these sets may *complement* each other (Podsakoff et al. 2012). (This new set may consist of just one manifest indicator or a latent indicator with multiple manifest indicators, as depicted in Table 3).

In most cases, however, indicators suffer from many sources of error. They often relate to various features of measurement common to the indicators, including the time and location of assessment (and thus mood, energy, mind-set, expectations, etc.), the language of questions, and the priming influence of previous content (Burton-Jones 2009; Podsakoff et al. 2003; Podsakoff et al. 2012; Tourangeau et al. 2000). They may also relate to a specific view of the construct, such as a cause of its effects, or the sum of its parts. For example, it may be unclear when a construct causes its effects, and when these effects are caused by alternative causes.

A way to alleviate this issue is to expand the model by specifying multiple *overlapping* indicators, where one is latent, and measured by the original indicators. The other(s), being either manifest or latent, should

be differentially sensitive to the sources of error, such that they overlap with the meaning of the construct. For example, self-report may be combined with peer-report; a single survey may be combined with the momentary assessment method; a construct as measured as a sum of its parts may be combined with an indicator that refers to the multiplication of its dimensions, etc.

New indicators should not be clearly inferior to the extant ones. Figure 3 illustrates a hypothetical example in which we know the quality and quantity of errors across five indicators. Two of them are clearly inferior, while three others are heterogeneous in their error, making them suitable for combination. (Note that in this error diagram, overlap is *undesirable*.) This may trigger a new round of evaluation and redesign, toward better confidence in validity.

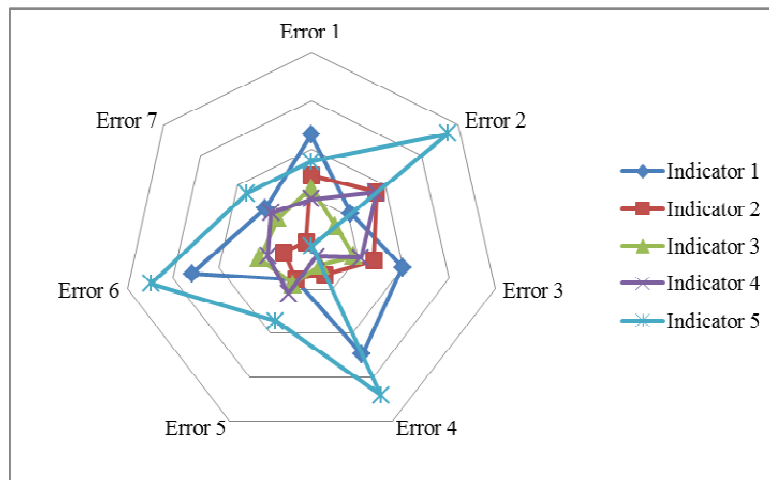


Figure 3: An Illustration of a Comparison of Indicators

Demonstration

To demonstrate these recommendations on developing measurement, I consider a specific example in the domain of IT addiction. With this example I illustrate various – not all – sources of error of a model of measurement, and emphasize how expanding it methodically can aid validity.

As part of an inquiry into the effects of smartphone addiction, suppose we are interested in how often smartphone users have a temptation to use their device to examine the effects of such addiction. That is, how often do they have an impulse to start using it while they should be doing other things? This can be called the frequency of the need to refrain from smartphone use, or ‘Need to Refrain’ for short. The objective is to capture the variance of this frequency across individuals, rather than a specific number of times for each individual.

An Initial Model

Following Netemeyer’s (2003) advice, I started by considering a simple single-indicator measurement model, asking each smartphone-using participant in one questionnaire at one point in time:

How often do you have a temptation to use your smartphone?

Never O O O O O O O Constantly

One cause of concern with this item is that individuals may have different conceptions of ‘temptation’. Is a temptation always a consciously experienced feeling of wanting or can a temptation be an automatic, habitual, and non-conscious impulse? Further, when are these impulses really temptations? Must there be some *feeling* that it is bad or wrong? Or is it also a temptation when it is rationalized or justified? These questions are about differences in conception across participants that may undermine the interpretation of records and therefore also the validity of research inferences. We can reduce these differences by providing a short description of what is meant by a temptation:

How often do you have a temptation to use your smartphone?

(A temptation is a conscious or subconscious impulse that you need to resist.)

Never Constantly

Model Expansion

Another concern with the initial model relates to the cognitive evaluation of the question. Will respondents be able to give an accurate answer? Certainly, the frequency of smartphone temptations will be less accessible than the frequency of meals on a day. Individuals differ in how they respond to questions which answers are less accessible (Tourangeau et al. 2000). They may be more accurate when they rely on systematic analysis than on gut feeling. Many guidelines recommend assisting such an analysis with more specific questions (Churchill 1979; Dillman 2000). We may generate indicators that refer to constituent parts: that is, we could cut up the substance of the construct.

This can be done from different angles, such as the goal the temptation conflicts with (e.g. *When you need to study, how often do you have an impulse to use your smartphone instead?*), the object of the temptation (e.g. *How often do you have a temptation to check for new text messages?*), the time of its occurrence (e.g. *In the morning, how often...*), its location (e.g. *At work, how often...*), the presence of others during the temptation (e.g. *When you're alone,...*), the mood in which the temptation occurs (e.g. *When you feel happy, ...*), etc. Thought experiments can help evaluate these options.

Operationally, these angles will vary in the ease with which their corresponding questions can be interpreted and answered. For example, answering a question about a temptation in concurrence with a certain mood state like worry or happiness seems intuitively harder than answering a question about a temptation in a certain location. Further, the 'goal' angle seems to generate relatively easy questions as it pertains to the meaning of the temptation directly, lending the questions coherence (Pinker 2014). Conceptually, not all angles can easily result in mutually exclusive and exhaustive sets of indicators, or indicators that are representative of such a set. Further, the weighting of different indicators is obvious for some angles, but not for others. 'Time' is easy to cut up, whereas the use of other angles requires additional information.

After deliberation, I opted for using conflicting goal as an angle of analysis. I specified three goal-specific items, namely to read, to write, and to listen, resulting in three questions: *When you need to [read, write, listen] how often...* I also included additional questions about the salience of these goals (i.e. *For your studies, how much do you need to [read, write, listen]*). While these extra questions provided a means to weight the part indicators, they also helped streamline the questionnaire, as participants were led to think about a goal first and then about an event given that goal. This lent further confidence in the conceptual and operational validity of the model.

I decided to expand one-item model rather than replace it, because the 'goal general' and an overall 'goal-specific' indicator overlap, and may be differentially sensitive to error. Specifically, the goal-general one could help correct for error due to the chosen goals not being representative of all conflicting goals.

Yet this rationale does not provide the information about how the two indicators can be combined mathematically. To evaluate various configurations, I ran an online pilot test, through Prolific Academic (n=151 English-speaking smartphone users of over 18 years of age). I included questions measuring constructs that were expected to correlate with the estimates, allowing tests of both internal and external validity. Specifically, I inserted a social desirability scale (Reynolds 1982), a self-control scale (Tangney et al. 2004), a question on how impulsive people thought they were, and questions related to smartphone use, such as how often they use different features, and how often they feel guilty about using it. I also asked participants to reflect on their experience of answering the questions, and report on the perceived difficulty and clarity of various parts of the questionnaire. This helped detect and locate any operational sources of error.

An Adjustment and a Second Expansion

While a low correlation with the social desirability scale suggested that responses to the items were only marginally contaminated by a tendency to provide social desirable rather than truthful answers, I re-

mained concerned with the interpretation of the item responses. Common to all items is a reliance on interpreting the question and accessing relevant memories during one session of filling out a questionnaire. The different feelings across participants and the content of their working memory may influence such processes, thereby introducing unwanted variability (Podsakoff et al. 2003). Further, the occurrence of temptations may leave few traces in memory (Hofmann et al. 2009). The degree to which it leaves traces depends on the circumstances of the impulse. This may also cause unwanted variability in the responses, because these circumstances may be systematically different across individuals.

I therefore adjusted the model and expanded it with one more indicator. The adjustment aimed to remove variability due to different psychological states across participants at the time of assessment. This was done through methodically specifying the items: I stipulated that participants filled out the questionnaire at one location, on one type of device, and after having performed tasks in an experiment. The questions were unchanged.

The expansion of the model consisted of an indicator based on reports at multiple times on the occurrence of smartphone temptations in last hour (this was based on a previous study on everyday temptations; Hofmann et al. 2012). As the added indicator captures the entire construct in a way that is different from the goal-specific and goal-general indicators, the indicators are overlapping. The added indicator should suffer less from errors due to memory (Hofmann et al. 2012), while it introduces an error that the others do not suffer from: error due to the sampling of the time of assessment. A combination of these indicators should thus attenuate the effect of these errors on the construct estimate.

Again, however, the conceptual basis for combining the indicators does inform one mathematical implementation. I could combine indicators in different sequences, as shown with three examples in Figure 4. In addition to these examples, it is possible to view the indicators as *reflective* in the sense that a certain habit manifests itself in the indicators, providing a common rationale for using confirmatory factor analysis. I leveraged this conceptual space by empirically comparing various configurations of the combined measurement model.

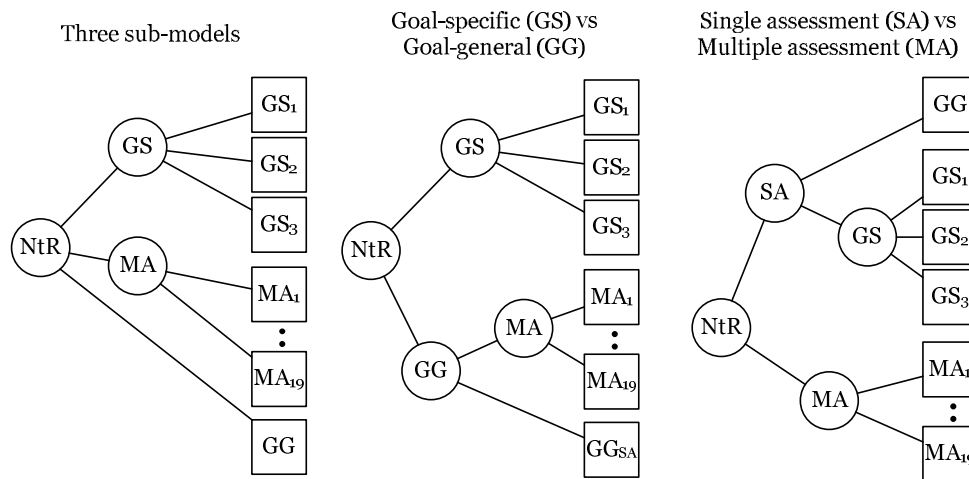


Figure 4. Three Model Structures with the Same Items

An Empirical Evaluation

I collected data once in a behavioral laboratory and across time through the experience sampling method (also called the momentary assessment method; Csikszentmihalyi and Larson 1987; Hektner et al. 2007). Student participants ($n=145$)¹ who used a smartphone were recruited to complete a series of questions and tasks in the lab. The questions corresponded to the goal-general and the goal-specific indicators, and

¹ A sample of 145 was thought to be sufficient for the analyses I planned to make, based on advice by Netemeyer et al. (1993).

to variables that were expected to correlate with the Need to Refrain. At the end of this session, I invited participants to join the experience sampling study: 87% accepted the invitation.

A week later, I started sending text messages to those who joined. Each contained a link to a 2-minute online questionnaire. One of its questions asked them whether they had had any smartphone temptation in the hour before they received the message. To each participant, I sent two messages daily at random moments at daytime over a period of two work weeks, totaling twenty the number of messages. The last message contained a link to a different questionnaire that was designed to detect an influence of the method on the constructs themselves and on the process of reporting on them. I offered a monetary reward for each questionnaire that was completed within 30 minutes after reception of the message, with a minimum of eleven timely responses. The indicator I added to the model was the portion of timely responses in which a participant indicated that a temptation had occurred.²

Figure 5 shows results from tests of internal and external validity. It provides correlations of construct estimates of Need to Refrain across twelve configurations of the model. To aid comparison, these configurations include both the items of the questionnaire separately (numbered 1 to 4) and combinations of items (numbered 5 to 12). Combinations were common factor scores, averages, or based on both averages and common factor scores. Some common factor scores were based on indicators with little internal consistency, as can be inferred from the relevant correlations reported in the figure. The scores were included in the analysis to aid comparison. Averages were based on standardized scores when scales were heterogeneous).

While an infinite number of combinations would have been consistent with the conceptual relationships, the twelve combinations were the easiest to interpret, and ranged widely in the distribution of weights across indicators, providing an impression of all possible results. The reported correlations across the 12 configurations are color-scaled, from red (lowest) to green (highest). The average correlation across configuration for each configuration ('average internal') is provided in the row before last. It suggests that single indicators perform worse than multiple indicators; the highest average is for configurations that rely on all indicators.

Further, for each configuration, correlations are provided with seven external variables, each of which was assessed during the session in the lab. Each of these variables was designed and coded such that conceptually, a higher positive correlation is expected. Each row of the external correlates is color-scaled separately to aid comparison of configurations. This also holds for the rows that report the average correlations at the bottom.

These results show a general tendency that combining more heterogeneous indicators improves the prediction of external correlates. Configurations 10 to 12 rely on a combination of all indicators and are the most predictive. Configuration 9 also relies on all indicators, but the common factor score relied only marginally on the momentary assessment indicator, explaining its lower external validity. Perhaps surprisingly, this multiple assessment indicator (5) was on average a better predictor of the (lab-assessed) external correlates compared to any combination of the lab-assessed indicators (1-4 and 6-8).

These findings underscore the danger of relying on a single view or a single method of capturing a construct. Had I used only a single questionnaire to measure Need to Refrain, I would have found lower correlations across all external correlates. This could mean the difference between support and lack of support for research hypotheses.

² To strike a balance between statistical power and avoiding error due to time-sampling, I used the SMS records of a participant when at least ten surveys were returned. Given the setup of the SMS study, there was no way to validate each response.

	1	2	3	4	5	6	7	8	9	10	11	12	
Internal configurations	1: Read item												
	2: Write item	0.65											
	3: Listen item	0.15	0.24										
	4: One goal general item	0.62	0.54	0.31									
	5: Multiple assessment (SMS)	0.14	0.11	0.22	0.22								
	6: Goal specific (average 1-3)	0.80	0.84	0.62	0.66	0.20							
	7: Goal specific (weighted average 1-3)	0.78	0.74	0.22	0.59	0.21	0.77						
	8: Common factor score (1-4)	0.92	0.84	0.30	0.81	0.18	0.91	0.82					
	9: Common factor score (1-5)	0.95	0.86	0.39	0.76	0.20	0.93	0.86	1.00				
	10: Three models (average 4-6)	0.73	0.67	0.55	0.83	0.61	0.83	0.73	0.85	0.84			
	11: Multiple vs Single (average 5, 8)	0.71	0.65	0.40	0.68	0.75	0.75	0.70	0.78	0.79	0.96		
	12: Specific vs General (average [average 4, 5], 6)	0.80	0.76	0.61	0.81	0.50	0.92	0.78	0.91	0.91	0.98	0.92	
External correlates	Temptations during Experiment	0.04	0.07	0.10	0.11	0.35	0.09	0.05	0.08	0.13	0.28	0.31	0.23
	Phone Use during Experiment	0.00	-0.01	0.05	0.06	0.45	0.02	0.00	0.02	0.12	0.30	0.36	0.23
	Guilt about use	0.30	0.24	0.08	0.38	0.23	0.28	0.26	0.35	0.33	0.42	0.38	0.38
	"Phone hurts Focus"	0.27	0.27	0.17	0.40	0.26	0.32	0.20	0.36	0.37	0.45	0.42	0.42
	Sleep Quality	0.08	0.18	0.07	0.15	0.10	0.15	0.22	0.15	0.21	0.23	0.21	0.23
	Impulsivity	0.10	0.21	0.10	0.26	-0.06	0.18	0.21	0.20	0.21	0.17	0.12	0.18
	Smartphone use	0.13	0.21	0.18	0.12	0.44	0.23	0.19	0.18	0.30	0.41	0.47	0.40
Overall average correlation	0.39	0.39	0.23	0.40	0.25	0.45	0.40	0.46	0.48	0.52	0.50	0.52	
Average internal	0.66	0.63	0.36	0.62	0.30	0.75	0.65	0.76	0.77	0.78	0.74	0.81	
Average external	0.09	0.12	0.08	0.16	0.19	0.13	0.13	0.14	0.17	0.23	0.24	0.21	

Figure 5. Internal and External Validity of Twelve Configurations

Discussion

Lack of validity is often a hidden and complicated problem. Error may stem from a wide range of factors, many of which are difficult to recognize and control for. While the extant measurement literature has provided guidelines on various relevant aspects such as handling specific types of error (Podsakoff et al. 2003; Podsakoff et al. 2012), and validating measurement (MacKenzie et al. 2011; Straub 1989), I have attempted to synthesize the literature to provide recommendations to develop valid measurement in an integral framework to allow for the selection of heterogeneous indicators.

These indicators may specify multiple perspectives from which a construct can be viewed. For example, many constructs can both be viewed as the sum of its parts or the cause of its effects. The differences in these perspectives are associated with different types of threats to validity, such that the combination of perspectives should provide fertile ground for valid measurement. Heterogeneous indicators could also refer to different methods of data collection. Such indicators are especially promising since errors common to indicators are often associated with specific aspects of the measurement method, such as the time of assessment, the questionnaire, or the order of items (Burton-Jones 2009; Drury and Farhoomand 1997; Podsakoff et al. 2012). Appending such measurement with other methods of inquiry could hold the promise of more valid measurement, as depicted in Figure 6.

This figure is meant to illustrate that the potential increase in validity by adopting heterogeneous indicators may vary – and what this implies in different types of studies. Sometimes, a single indicator is sufficient. In most cases of measuring sex, age, marital status, occupation, and consent, for example, little can be gained. In other circumstances, having multiple, homogeneous indicators, like multiple questions in a questionnaire that only differ syntactically, may be sufficient for the purposes of a study.

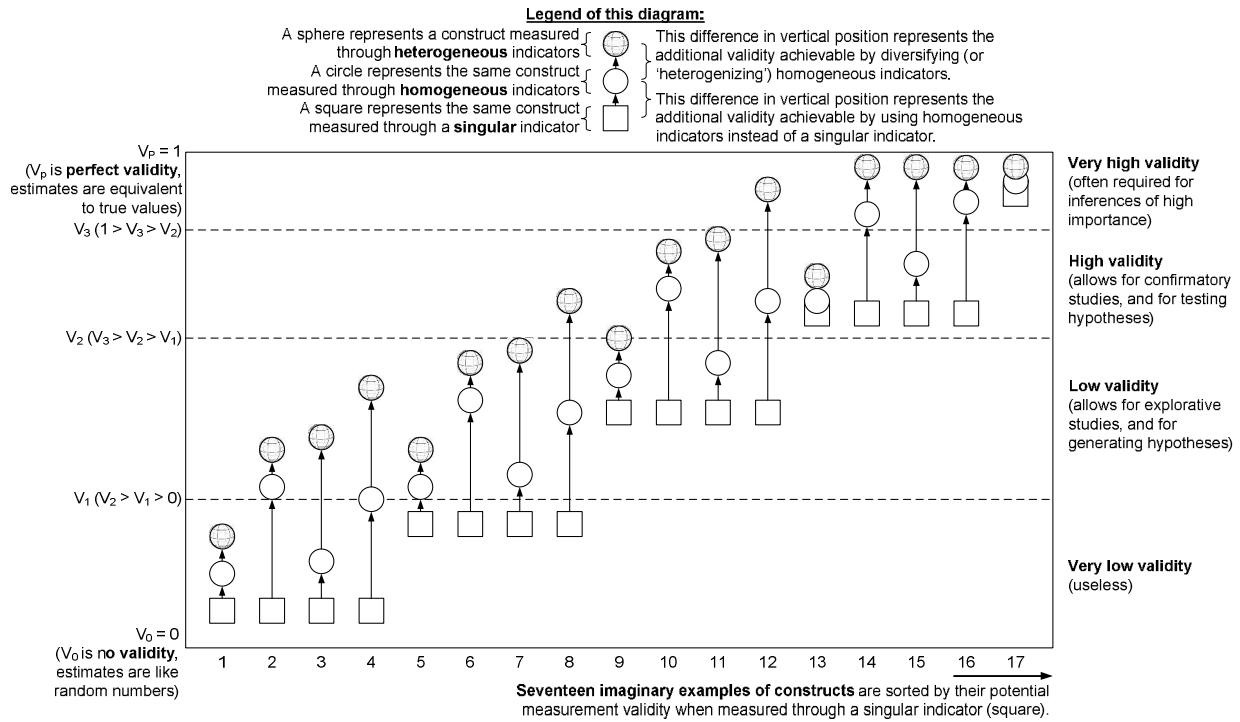


Figure 6. Potential Added Validity by Diversifying Indicators

Adopting heterogeneous indicators is, however, worth considering in many circumstances. In general, more valid measurement can prevent mistakes in drawing research conclusions (Viswanathan 2005). Such mistakes are expensive when they lead a stream of research down a blind alley. Multiple methods are more likely to prevent wrong inferences when research conclusions are more sensitive to measurement validity, such as when studying small effects or using small samples. That is, better measurement validity can – to some extent – compensate for lack of statistical power. Further, using multiple methods should pay off more when even the best measurement model of a construct suffers from much error.

Various technologies have helped lower operating costs of measurement methods. For instance, the internet and mobile devices have made it easier to measure through the momentary assessment method. This approach complements more traditional approaches, being less sensitive to errors specific to location and time (Hektner et al. 2007; Podsakoff et al. 2012). Innovations in digital payments and location-based services will reduce the administrative burden of such measurement methods, and increase its usefulness.

One limitation of more heterogeneous measurement is that it may complicate the standardization of measurement. While standardization generally helps in comparing multiple studies, the premise of meaningful comparisons is that constructs are measured with sufficient validity. The adoption of more heterogeneous indicators can help ensure this premise. I would thus argue that this adoption aides comparison.

Future research could shed more light on the exploitation of conceptual leeway in the combination of indicators, given various model structures, and expectations on how different sources of error threaten parts of this structure.

Error can threaten the validity of measurement in a plethora of ways, especially in survey research. Recommended procedures and techniques to control for these errors are limited because of their underlying assumptions. They may assume a single view of a construct or a single method of capturing it. I hope that by contextualizing and typifying the ways in which a construct can be tied to indicators this paper will lead researchers come up with heterogeneous indicators to measure their construct of interest with confidence in validity.

Acknowledgements

I am grateful to my doctoral advisor, Professor Ali Farhoomand, who has led me to dive deeper into the topic of measurement, and my collaborator, Professor Israr Qureshi, for his continuous help in improving the manuscript. I also thank Professor Jason Thatcher, Dr. Aguirre-Urreta, and anonymous reviewers for their insightful comments on an earlier version of this manuscript.

References

- Aguirre-Urreta, M.I., and Marakas, G.M. 2012. "Revisiting Bias Due to Construct Misspecification: Different Results from Considering Coefficients in Standardized Form," *MIS Quarterly* (36:1), pp 123-138.
- Aguirre-Urreta, M.I., and Marakas, G.M. 2013. "Research Note—Partial Least Squares and Models with Formatively Specified Endogenous Constructs: A Cautionary Note," *Information systems research* (25:4), pp 761-778.
- Aguirre-Urreta, M.I., and Marakas, G.M. 2014. "A Rejoinder to Rigdon Et Al.(2014)," *Information Systems Research* (25:4), pp 785-788.
- Barki, H. 2008. "Thar's Gold in Them Thar Constructs," *ACM SIGMIS Database* (39:3), pp 9-20.
- Bergkvist, L., and Rossiter, J.R. 2007. "The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs," *Journal of marketing research* (44:2), pp 175-184.
- Blalock, H.M. 1964. *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press.
- Bollen, K. 1989. "Structural Equations with Latent Variables." New York: Wiley.
- Bollen, K., and Lennox, R. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective," *Psychological Bulletin* (110:2), p 305.
- Bollen, K.A., and Bauldry, S. 2011. "Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates," *Psychological methods* (16:3), p 265.
- Boudreau, M.-C., Gefen, D., and Straub, D.W. 2001. "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Quarterly* (25:1), pp 1-16.
- Burton-Jones, A. 2009. "Minimizing Method Bias through Programmatic Research," *MIS Quarterly* (33:3), pp 445-471.
- Carlson, L., and Grossbart, S. 1988. "Parental Style and Consumer Socialization of Children," *Journal of Consumer Research* (15:1), pp 77-94.
- Cenfetelli, R.T., and Bassellier, G. 2009. "Interpretation of Formative Measurement in Information Systems Research," *MIS Quarterly* (33:4), pp 689-707.
- Chin, W.W., Thatcher, J.B., and Wright, R.T. 2012. "Assessing Common Method Bias: Problems with the Ulmc Technique," *MIS Quarterly* (36:3), pp 1003-1019.
- Churchill, G.A. 1979. "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of marketing research* (16:1), pp 64-73.
- Clark, L.A., and Watson, D. 1995. "Constructing Validity: Basic Issues in Objective Scale Development," *Psychological Assessment* (7:3), p 309.
- Cronbach, L.J. 1951. "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika* (16:3), pp 297-334.
- Csikszentmihalyi, M., and Larson, R. 1987. "Validity and Reliability of the Experience-Sampling Method," *The Journal of nervous and mental disease* (175:9), pp 526-536.
- DeVellis, R.F. 2003. *Scale Development: Theory and Applications*. Thousand Oaks, California: Sage Publications.
- Diamantopoulos, A. 2011. "Incorporating Formative Measures into Covariance-Based Structural Equation Models," *MIS quarterly* (35:2), pp 335-358.
- Diamantopoulos, A., Riefler, P., and Roth, K.P. 2008. "Advancing Formative Measurement Models," *Journal of Business Research* (61:12), pp 1203-1218.
- Diamantopoulos, A., and Siguaw, J.A. 2006. "Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration," *British Journal of Management* (17:4), pp 263-282.
- Diamantopoulos, A., and Temme, D. 2013. "Mimic Models, Formative Indicators and the Joys of Research," *AMS review* (3:3), pp 160-170.
- Diamantopoulos, A., and Winklhofer, H.M. 2001. "Index Construction with Formative Indicators: An Alternative to Scale Development," *Journal of marketing research* (38:2), pp 269-277.
- Dillman, D.A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. Wiley New York.

- Drury, D.H., and Farhoomand, A. 1997. "Improving Management Information Systems Research: Question Order Effects in Surveys," *Information Systems Journal* (7:3), pp 241-251.
- Edwards, J.R. 2001. "Multidimensional Constructs in Organizational Behavior Research: An Integrative Analytical Framework," *Organizational research methods* (4:2), pp 144-192.
- Edwards, J.R. 2011. "The Fallacy of Formative Measurement," *Organizational research methods* (14:2), pp 370-388.
- Edwards, J.R., and Bagozzi, R.P. 2000. "On the Nature and Direction of Relationships between Constructs and Measures," *Psychological methods* (5:2), p 155.
- Epstein, S. 1983. "Aggregation and Beyond: Some Basic Issues on the Prediction of Behavior," *Journal of personality* (51:3), pp 360-392.
- Fornell, C., and Larcker, D.F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18:1), pp 39-50.
- Goertz, G. 2006. *Social Science Concepts: A User's Guide*. Princeton University Press.
- Harman, H.H. 1976. *Modern Factor Analysis*. Chicago: University of Chicago Press.
- Haynes, S.N., Richard, D., and Kubany, E.S. 1995. "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods," *Psychological Assessment* (7:3), p 238.
- Hektner, J.M., Schmidt, J.A., and Csikszentmihalyi, M. 2007. *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage.
- Hofmann, W., Baumeister, R.F., Förster, G., and Vohs, K.D. 2012. "Everyday Temptations: An Experience Sampling Study of Desire, Conflict, and Self-Control," *Journal of Personality and Social Psychology* (102:6), pp 1318-1335.
- Hofmann, W., Friese, M., and Strack, F. 2009. "Impulse and Self-Control from a Dual-Systems Perspective," *Perspectives on Psychological Science* (4:2), pp 162-176.
- Howell, R.D., Breivik, E., and Wilcox, J.B. 2013. "Formative Measurement: A Critical Perspective," *ACM SIGMIS Database* (44:4), pp 44-55.
- Jarvis, C.B., MacKenzie, S.B., and Podsakoff, P.M. 2012. "The Negative Consequences of Measurement Model Misspecification: A Response to Aguirre-Urreta and Marakas," *MIS Quarterly* (36:1), pp 139-146.
- Kaplan, A. 1964. *The Conduct of Inquiry: Methodology for Behavioral Science*. Chandler.
- Kim, G., Shin, B., and Grover, V. 2010. "Investigating Two Contradictory Views of Formative Measurement in Information Systems Research," *MIS Quarterly* (34:2), pp 345-365.
- King, W.R., Liu, C.Z., Haney, M.H., and He, J. 2007. "Method Effects in Is Survey Research: An Assessment and Recommendations," *Communications of the Association for Information Systems* (20:1), p 30.
- Klein, G., Jiang, J.J., and Cheney, P. 2009. "Resolving Difference Score Issues in Information Systems Research," *MIS quarterly*, pp 811-826.
- Kumar, N., Stern, L.W., and Anderson, J.C. 1993. "Conducting Interorganizational Research Using Key Informants," *Academy of Management Journal* (36:6), pp 1633-1651.
- Law, K.S., Wong, C.-S., and Mobley, W.M. 1998. "Toward a Taxonomy of Multidimensional Constructs," *Academy of Management Review* (23:4), pp 741-755.
- Lee, N., Cadogan, J.W., and Chamberlain, L. 2013. "The Mimic Model and Formative Variables: Problems and Solutions," *AMS review* (3:1), pp 3-17.
- Loevinger, J. 1957. "Objective Tests as Instruments of Psychological Theory: Monograph Supplement 9," *Psychological reports* (3:3), pp 635-694.
- MacKenzie, S.B. 2003. "The Dangers of Poor Construct Conceptualization," *Journal of the Academy of Marketing Science* (31:3), pp 323-326.
- MacKenzie, S.B., Podsakoff, P.M., and Podsakoff, N.P. 2011. "Construct Measurement and Validation Procedures in Mis and Behavioral Research: Integrating New and Existing Techniques," *MIS Quarterly* (35:2), pp 293-334.
- Markus, K.A., and Borsboom, D. 2013. *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. Routledge.
- Meredith, W. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance," *Psychometrika* (58:4), pp 525-543.
- Nederhof, A.J. 1985. "Methods of Coping with Social Desirability Bias: A Review," *European Journal of Social Psychology* (15:3), pp 263-280.
- Netemeyer, R.G., Bearden, W.O., and Sharma, S. 2003. *Scaling Procedures: Issues and Applications*. Sage.
- Nunnally, J.C., and Bernstein, I.H. 1994. *Psychometric Theory*, (3 ed.). New York, NY: McGraw-Hill.

- Ortiz de Guinea, A., and Webster, J. 2013. "An Investigation of Information Systems Use Patterns: Technological Events as Triggers, the Effect of Time, and Consequences for Performance," *MIS Quarterly* (37:4), pp 1165-1188.
- Os, J., Delespaul, P., Wigman, J., Myin-Germeys, I., and Wichers, M. 2013. "Beyond Dsm and Icd: Introducing "Precision Diagnosis" for Psychiatry Using Momentary Assessment Technology," *World Psychiatry* (12:2), pp 113-117.
- Petter, S., Rai, A., and Straub, D. 2012. "The Critical Importance of Construct Measurement Specification: A Response to Aguirre-Urreta and Marakas," *MIS Quarterly* (36:1), pp 147-155.
- Petter, S., Straub, D., and Rai, A. 2007. "Specifying Formative Constructs in Information Systems Research," *MIS Quarterly* (31:4), pp 623-656.
- Pinker, S. 2014. *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. Penguin.
- Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., and Podsakoff, N.P. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology* (88:5), pp 879-903.
- Podsakoff, P.M., MacKenzie, S.B., and Podsakoff, N.P. 2012. "Sources of Method Bias in Social Science Research and Recommendations on How to Control It," *Annual Review of Psychology* (63), pp 539-569.
- Polites, G.L., Roberts, N., and Thatcher, J. 2012. "Conceptualizing Models Using Multidimensional Constructs: A Review and Guidelines for Their Use," *European Journal of Information Systems* (21:1), pp 22-48.
- Reynolds, W.M. 1982. "Development of Reliable and Valid Short Forms of the Marlowe-Crowne Social Desirability Scale," *Journal of Clinical Psychology* (38:1), pp 119-125.
- Rigdon, E.E. 2013. "Lee, Cadogan, and Chamberlain: An Excellent Point... But What About That Iceberg?," *AMS review* (3:1), pp 24-29.
- Rigdon, E.E., Becker, J.-M., Rai, A., Ringle, C.M., Diamantopoulos, A., Karahanna, E., Straub, D., and Dijkstra, T.K. 2014. "Conflating Antecedents and Formative Indicators: A Comment on Aguirre-Urreta and Marakas," *Information Systems Research* (25:4), pp 780-784.
- Ringle, C.M., Sarstedt, M., and Straub, D. 2012. "A Critical Look at the Use of PLS-Sem in MIS Quarterly," *MIS Quarterly (MISQ)* (36:1).
- Segars, A.H., and Grover, V. 1998. "Strategic Information Systems Planning Success: An Investigation of the Construct and Its Measurement," *MIS Quarterly* (22:2), pp 139-163.
- Sharma, R., Yetton, P., and Crawford, J. 2009. "Estimating the Effect of Common Method Variance: The Method—Method Pair Technique with an Illustration from Tam Research," *MIS Quarterly* (33:3), pp 473-490.
- Spector, P.E. 1992. *Summated Rating Scale Construction: An Introduction*. Sage.
- Spector, P.E. 2006. "Method Variance in Organizational Research Truth or Urban Legend?," *Organizational research methods* (9:2), pp 221-232.
- Straub, D.W. 1989. "Validating Instruments in MIS Research," *MIS quarterly* (13:2), pp 147-169.
- Tangney, J.P., Baumeister, R.F., and Boone, A.L. 2004. "High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success," *Journal of Personality* (72:2), pp 271-324.
- Tourangeau, R., Rips, L.J., and Rasinski, K. 2000. *The Psychology of Survey Response*. Cambridge University Press.
- Van de Ven, A.H. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research: A Guide for Organizational and Social Research*. Oxford University Press.
- Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F.D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp 425-478.
- Viswanathan, M. 2005. *Measurement Error and Research Design*. Sage.
- Wetzels, M., Odekerken-Schröder, G., and Van Oppen, C. 2009. "Using PLS Path Modeling for Assessing Hierarchical Construct Models: Guidelines and Empirical Illustration," *MIS Quarterly* (33:1), pp 177-195.
- Woszczyński, A.B., and Whitman, M.E. 2004. "The Problem of Common Method Variance in IS Research," in: *The Handbook of Information Systems Research*, A.B. Woszczyński and M.E. Whitman (eds.). Hershey, PA: Idea Group Inc., pp. 66-77.