Anticipating, avoiding, and alleviating measurement error: A synthesis of the literature with practical recommendations

Sander Zwanenburg
University of Otago
sander.zwanenburg@otago.ac.nz

Israr Qureshi
ANU College of Business and Economics

Abstract

Researchers' ability to draw inferences from their empirical work hinges on the degree of measurement error. The literature in Information Systems and other behavioural disciplines describes a plethora of sources of error. While it helps researchers deal with them when taking specific steps in the measurement process, like modelling constructs, developing instruments, collecting data, and analysing data, it does not provide an overall guide to help them prevent and deal with measurement error. This paper presents a synthesis of the insights in the literature through a decomposition of the logic of measurement. It shows how researchers can classify sources of error, evaluate their impact, and refine their measurement plans, in terms of specific steps or overall measurement approaches. We hope this will aid researchers in anticipating, avoiding, and alleviating error in measurement, and in drawing valid research conclusions.

Keywords Measurement, construct, indicator, model, operationalization.

1 Introduction

The ability to draw valid inferences from empirical research in Information Systems and other behavioural disciplines hinges on the degree of measurement error (Cote & Buckley, 1988; MacKenzie, Podsakoff, & Podsakoff, 2011). Decades of methodological studies on measurement error have helped researchers and reviewers understand and recognize a variety of sources of such error (Podsakoff, MacKenzie, & Podsakoff, 2012). They have also helped in the development of measurement guidelines (Dillman, 2000; Lewis, Templeton, & Byrd, 2005; MacKenzie et al., 2011).

However, measurement problems persist. In one Information Systems journal, for example, five of eight main obstacles that are frequent causes of desk rejects are (1) failures to recognize a lack of construct clarity, (2) common method bias, (3) formative constructs, (4) issues with self-report data, and (5) issues with data-analytic techniques (Gregor & Klein, 2014). Existing measurement practices "fail to address the full landscape of measurement issues and fail to prioritize the fundamental aspects" (p451, Burton-Jones & Lee, 2017).

More broadly, the need for more rigor in methodology has been underscored in the establishment of the AIS Transactions on Replication Research, and in the frequent failures to reproduce findings from behavioural research (Open Science Collaboration, 2012, 2015; Servick, 2018).

To help researchers measure well and draw valid conclusions, we see an opportunity in enhancing the provision of guidelines on anticipating, avoiding, and alleviating error. While

the huge number of studies on specific sources of measurement error, such as social desirability, common method, or violations of statistical assumptions, has helped us understand these sources, they may leave many researchers confused about what needs to be done about all of these given a measurement task (Bagozzi, 2011; Burton-Jones & Lee, 2017; MacKenzie et al., 2011; Spector, 1992). Arguably, researchers face "an almost unworkable number of tests to comply with" (Burton-Jones & Lee, 2017, p451) and "there are so many issues to consider … that we might be apt to throw up our hands in frustration" (Bagozzi, 2011, p288).

The literature that considers all measurement error, on the other hand, treat it either as a whole or unpacks it into abstract components, such as bias versus random error (Cote & Buckley, 1988), and item-level versus construct-level error (MacKenzie et al., 2011). Being abstract and evaluated after data is collected, they do not always lend themselves to draw concrete implications for improving measurement plans (Zyphur & Pierides, 2017). Compounding this issue, many guidelines that help researchers develop measurement, carry often implicit methodological assumptions (Bollen & Lennox, 1991; DeVellis, 2003; Lewis et al., 2005; MacKenzie et al., 2011). Common assumptions are that measurement relies on reflective models, self-report questionnaires, and on one-off assessments. Inadvertently, these assumptions may encourage uncritical adoption of ready-made formulas (Zyphur & Pierides, 2017), and prevent researchers from considering alternative methods or models of measurement, which could be valuable substitutes or complements (Zwanenburg, 2015).

We therefore believe that a comprehensive classification of the sources of measurement error and their potential remedies will help researchers anticipate, avoid, and alleviate measurement error through revisions of measurement plans. In this paper, we take initial steps toward this objective. Based on the literature on sources of measurement error, we introduce a framework to classify these, and present the classification. We then discuss how an evaluation of their potential impact can inform decisions to improve plans of measurement. This can apply to a wide range of measurement approaches, e.g. the use of reflective, formative, and other measurement models; questionnaire surveys and other data collection approaches; qualitative and quantitative measurement variables. Thus, we hope this will help many researchers systematically evaluate potential measurement error, and identify appropriate improvements to avoid or alleviate their impact.

2 The Meaning and Premise of Measurement Error

Across the wide range of contexts of measurement, the objective of measurement is to obtain estimates of a construct that fit the meaning of that construct. This fit is the logical basis for drawing inferences, such as research conclusions (Burton-Jones, 2009) and is known as the validity of measurement (Markus & Borsboom, 2013; Nunnally & Bernstein, 1994; O'Leary-Kelly & Vokurka, 1998; Peter, 1981).[1] Error of measurement is the inverse: the gap between what is to be measured and what is actually measured (Nunnally & Bernstein, 1994).

[1] This differs from what is termed 'construct validity,' a property of test score interpretations (see e.g. Borsboom et al. 2009, Cronbach 1989). It also differs from validity as 'the lack of *systematic* error' (e.g. Carmines and Zeller, 1979; Adcock and Collier 2001), as complementary to 'reliability' as the lack of *random* error. In our definition, reliability is a form of validity. Other forms, like content validity, cross validity, face validity, refer to positive findings from specific tests that can indicate problems with

This definition is consistent with various classical theories of measurement error, viewing error as the difference between the true value and the observed value (Nunnally & Bernstein, 1994). True value here refers to the value that corresponds to the meaning of a construct when fully defined and applied to an instantiation.[2] Like its estimates, it can be categorical, continuous, or of another type. Note that *estimates* of error, sometimes confusingly called 'error' or 'standard error', can deviate from actual error because of violations to the assumptions underlying the methods of their estimation, such as linearity and independence of indicators.

What is error in a measurement thus depends on the meaning of the construct of that measurement. A lack of clarity prevents the researcher from evaluating error and thus the validity of measurement and the broader research conclusions. While a clear meaning of the construct is a straightforward premise to the evaluation of error, ambiguity is a common problem (Gregor & Klein, 2014; MacKenzie et al., 2011). Indeed, as DeVellis (1991, p51) noted: *many researchers think they have a clear idea of what they wish to measure, only to find out that their ideas are more vague than they thought. Frequently, this realization occurs after considerable effort has been invested in generating items and collecting data—a time when changes are far more costly than if discovered at the outset of the process*.

The meaning of a construct is clear when (a) the instantiations of the construct are clear, (b) each instantiation can only have one true value, and (c) the construct is embedded in a framework of other constructs.

2.1 Clear Instantiations

Instantiations of a construct can refer to entities or relationships between them (Burton-Jones & Lee, 2017).[3] For example, an app's usefulness can be thought of as an attribute of that app but *perceptions* of its usefulness are attributes of the relationship between this object and its observer (Gregor & Klein, 2014). The difference is critical, and choosing one over the other can carry many implications for the design of measurement. For example, if the target of measurement is the app's usefulness and multiple perceptions are a meansof accessing that, how are different perceptions combined in a measurement model?

Another important consideration in clarifying the instantiations of a construct is their relation to time. When targets of measurement are not tied to a particular moment or event, they can be subject to change over time. This can be slow or rapid, continuous or incidental: an app's user might 'get the hang of it', or might become frustrated with a bad update. One instantiation

validity based on the domain of the construct, the sample, or the inspection of measurement respectively.

[2] True values are sometimes called true scores, which operationalists see as outcomes of measurement processes themselves (Nunnally & Bernstein, 1994). This view is problematic because it detaches the concept of validity from the meaning of constructs (Markus & Borsboom, 2013).

[3] Some authors use the terms object and entities interchangeably. Here, the word *object* is used in relation to the measurement target itself and *entity* in relation to possible referents of it, as implied by the meaning of a construct and its instantiations. The object of a measurement can also refer to a relationship between entities.

of it can be a selection from multiple moments, or an aggregate.[4] Again, the difference can be critical. In sum, to evaluate measurement validity, it must be clear what the instantiations of a construct refer to structurally and temporally.

2.2 One Instantiation, One Value

A common issue with constructs is having multiple possible true values per instantiation, often due to multiple interpretations. An example is the construct of *smartphone use frequency*. Assuming a questionnaire is a method of choice, one might try measure it by asking, "*how often do you use your smartphone?*" with answer options ranging from *never* to *very often*. But what is *often*? A respondent may evaluate this through comparisons over time, comparisons with perceptions of their peers' behaviour, or with perceived domestic, organisational, or societal norms (Tourangeau, Rips, & Rasinski, 2000; Zyphur & Pierides, 2017). Different evaluations may correspond to different true values, introducing ambiguity and impeding the evaluation of validity. They may be prevented through a clarification, ultimately based on previous work, standards in the literature, nomological networks, and/or the specific purposes of the measurement (Zyphur & Pierides, 2017). For example, if for the specific purposes of the measurement it is desirable to have a social-relative measure, one can refine the definition of the construct and the operations of its measurement items to prevent this ambiguity (e.g. *Compared to your peers, how often do you use your smartphone?*).

2.3 Interrelated Constructs

Elucidating the relationships between related constructs will help identify ways to define a construct and measure it. Such relations may be deterministic or probabilistic. When they are known, they constitute a construct's 'internal theory' (Goertz, 2006; MacKenzie et al., 2011). Some relations might be the object of a study. They may also be part of or implied by the definition of the construct, such that measurement of the related concepts can be part of the measurement of the focal construct.

For example, someone's usage of a device will be equivalent to the usage of the device itself, which may be captured in automatic logs, minus the usage of the device's other users. Potentially, capturing these related concepts to measure the construct may produce better validity than a direct approach. Thus, an evaluation of related constructs not only aids the clarity of the construct, it can also identify ways to measure it.

In sum, identifying the instantiations, the multiplicity of their true values, and the relations with other constructs will be instrumental in anchoring this meaning conceptually and satisfying the premise of anticipating, avoiding, and alleviating error.

3 A Classification of Sources of Error

Sources of error undermine the logic of measurement in different ways and at different places. It may stem from random processes, affecting each data point separately, or systemic ones, affecting one or multiple arrays of data points, within or across data sets (Cote & Buckley, 1988). It may be attributed to a variety of aspects of a measurement, such as an instrument, instructions of a questionnaire, a question's wording or content, a response scale, a sequence

[4] In this paper, selecting and sampling instantiations to measure are important study design considerations but are not treated in this paper. While unmeasured instantiations are of concern to drawing research inferences, they are outside of the scope of measurement proper.

of questions, a trait or state of a respondent, a measurement model, an overall method of collecting data, and assumptions of estimation techniques. Error may be attributable to an interplay of factors associated with these aspects.

To allow researchers to systematically identify and evaluate potential sources of error that threaten the validity of their measurement, we first consider the logic of measurement in general and decompose it into conceptual, operational, and inferential components. We then illustrate how this logic can be further unpacked for particular instances of measurement. In the last Subsection, we will illustrate how this allows for a tailored classification of sources of error.

3.1 Decomposing the Logic of Measurement

Let us consider a single measurement instantiation such as the usefulness of a new app, using a non-zero set of indicators, which can be based, for example, on one or multiple questions, assessments, informants, or other sources of information. Logically, the measurement task is to use these indicators to connect the meaning of this construct's instantiation, i.e. an unobserved true value, with an estimate value. Unpacking this logical relationship can help classify error, as illustrated in Figure 1.

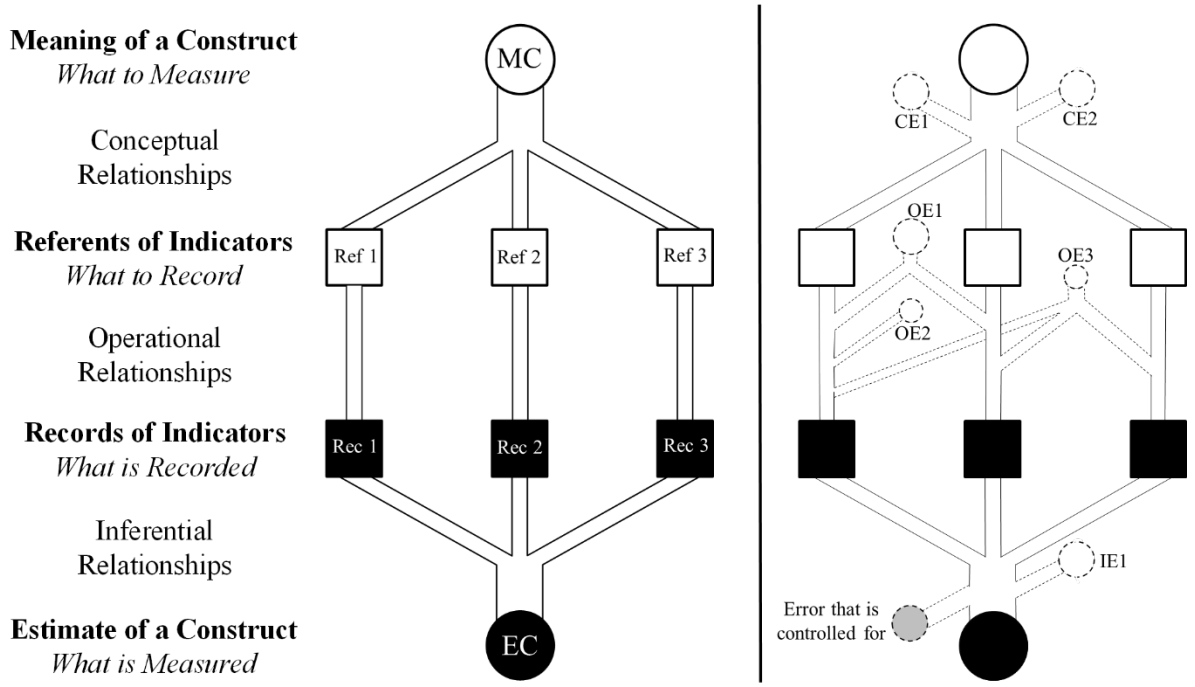

Figure 1: Left: A decomposition of the logical relationship of measurement into conceptual, operational, and inferential relationships, using three indicators. Right: While some error may be controlled for, various sources of error can affect the conceptual, operational, and inferential relationships (marked CE, OE, and IE for Conceptual, Operational, and Inferential error). This figure distinguishes between the targets of a measurement (i.e. what is to be captured) in white shapes and the actual data (i.e. what is actually captured) in black shapes.

At a generic level, we can decompose the logical relationship into conceptual, operational, and inferential ones, as shown in the left panel. In the diagram, the meaning of the construct on top is translated into referents of three indicators, which are operated to yield three records,

which are then combined to infer an estimate of the construct. As visualised in the right panel, sources of error can introduce interference to this logical flow at each of these steps, some of which can be removed though data analytical techniques.

3.1.1 Conceptual Relationships

Here, a conceptual relationship refers to the relation between the meaning of a construct and what an indicator is to record, i.e. its referent. This referent may be identical to the meaning of the construct, or it may be a part, cause, effect, or manifestation of it: anything that stands in some relation to the construct (Law, Wong, & Mobley, 1998; MacKenzie et al., 2011; Polites, Roberts, & Thatcher, 2012). For example, the Likert-type item "Learning to operate the Web site is easy" carries meaning that can be seen as identical to or a manifestation of the meaning of the construct, the ease of use of the same web site (Gefen, Karahanna, & Straub, 2003). When an item would refer to ease of use in *navigating* the web site, this can be seen as referring to part of the meaning of the construct. Conceptual relationships can be compounded (Edwards & Bagozzi, 2000). Such compounded relationships are sometimes explicitly modelled using sub-constructs (Law et al., 1998; Polites et al., 2012).

When accurate knowledge of the referents perfectly inform the true value of the construct, there is no error at this conceptual level. For example, if we know how much someone uses the web per weekday and in the weekend we can infer how much this person uses the web during the week, without error. Conceptual relationships are erroneous, however, when referents of indicators relate to the meaning of the construct in a non-deterministic way. For instance, measuring extraversion with a question about the frequency of going to parties relies on the idea that more extraverted people *tend* to go to parties more. Error may also relate to the referents of a set of indicators. This set may lack content validity, such as when they comprise a list that lacks mutual exclusivity or completeness, or when they include extraneous items (Haynes, Richard, & Kubany, 1995; Lewis et al., 2005; Messick, 1989).

3.1.2 Operational Relationships

An operational relationship refers to the relation between the referent of an indicator and its record, what is actually recorded. Error at this level stems from the actual physical and psychological processes that influence the generation of a datum. For question and answer-based measurement, for example, these include capturing the referent of an indicator in a question, expressing that question, hearing or reading it, interpreting it, evaluating it, responding to it, and capturing that response in a record (Dillman, 2000). Much error in question-based measurement stem from these processes (Podsakoff et al., 2012; Tourangeau et al., 2000). As indicated in the right panel of Figure 1, a source of error can introduce interference in multiple operations. For example, social desirability error or error stemming from an unclear questionnaire introduction can affect the validity of some or all of a construct's indicators.

3.1.3 Inferential Relationships

An inferential relationship refers to the relation between the records of the indicators and the estimate of a construct, as captured in mathematical and logical equations and operations. Inferential relationships can be modelled additively, multiplicatively, or otherwise, based on reflective, formative, and other models (MacKenzie et al., 2011; Mellenbergh, 1994; Polites et al., 2012). This is based on the conceptual and operational relationships including an understanding of how sources of error could affect them. For example, factor analytical

approaches typically assume that random error has affected each record, following a normal distribution with mean zero and a standard deviation that is estimated (Bollen & Lennox, 1991; Nunnally & Bernstein, 1994). Hence, these procedures can control for error to the degree it agrees with these assumptions. When these inferences are based on invalid assumptions or are erroneously implemented, they may fail to properly control for error and can introduce new error in the measurement process (Rigdon, 2012). For example, a factor model may be mis-specified or unidentifiable leading to false conclusions (Aguirre-Urreta & Marakas, 2012; Anderson & Gerbing, 1988; Jarvis, MacKenzie, & Podsakoff, 2003).

Together, the conceptual, operational, and inferential components of the logic of measurement constitute a generic error classification framework. It is complete and non-exclusive in that any source of measurement error can be associated with a shortcoming in these components. It can be used to structure and systematize the evaluation of sources of error regardless of the form or method of measurement, as no methodological assumptions have been made. However, because of this, these relationships can still be both abstract and complex. Unpacking them further for specific instances of measurement can shed further light on the sources of error that can threaten their validity.

3.2 Identifying Specific Logical Links

The conceptual, operational, and inferential relationships can be abstract and complex as they may constitute a chain of logical links (Law et al., 1998; Polites et al., 2012; Tourangeau et al., 2000). These more elementary links may connect one idea to one or multiple others, such as a concept to a proxy, an abstraction to its manifestations, a whole to its parts, a cause to one or multiple effects. Elucidating these links will ease the identification of sources of errors.

As an extreme example, consider capturing the ease of use of an app with a single estimate by using four-dimensional indicator data involving (1) multiple parts (e.g. functions) of that app, (2) multiple users (who act as informants), (3) multiple questions, and (4) multiple times of assessment. Here, four consecutive one-to-multiple links can bridge this zero-dimensional construct (i.e. its estimate is a single point or a zero-dimensional array[5]) with the 4D indicators at the conceptual level, as illustrated in Figure 2. Further consecutive links unpack the causal chain of the measurement operations, from presenting a specific question to recording a response.[6] At each consecutive link, we can identify potential sources of error by asking specific questions, like those listed on the right hand side of the Figure. Thus, decomposing the logic of a measurement into elementary logical can help researchers obtain a tailored error classification framework.

[5] The word 'dimension' is used here in the context of a dataset, where aspects or facets of a construct may be positions on one dimension (e.g. its composition); elsewhere these aspects or facets may be called dimensions themselves (Law et al., 1998; Polites et al., 2012).

[6] In this Figure, the inferential relationship is not further decomposed. While some of its steps can be modelled as sequential logical links, such as data cleaning and dealing with missing data, its estimation step is often iterative in nature.

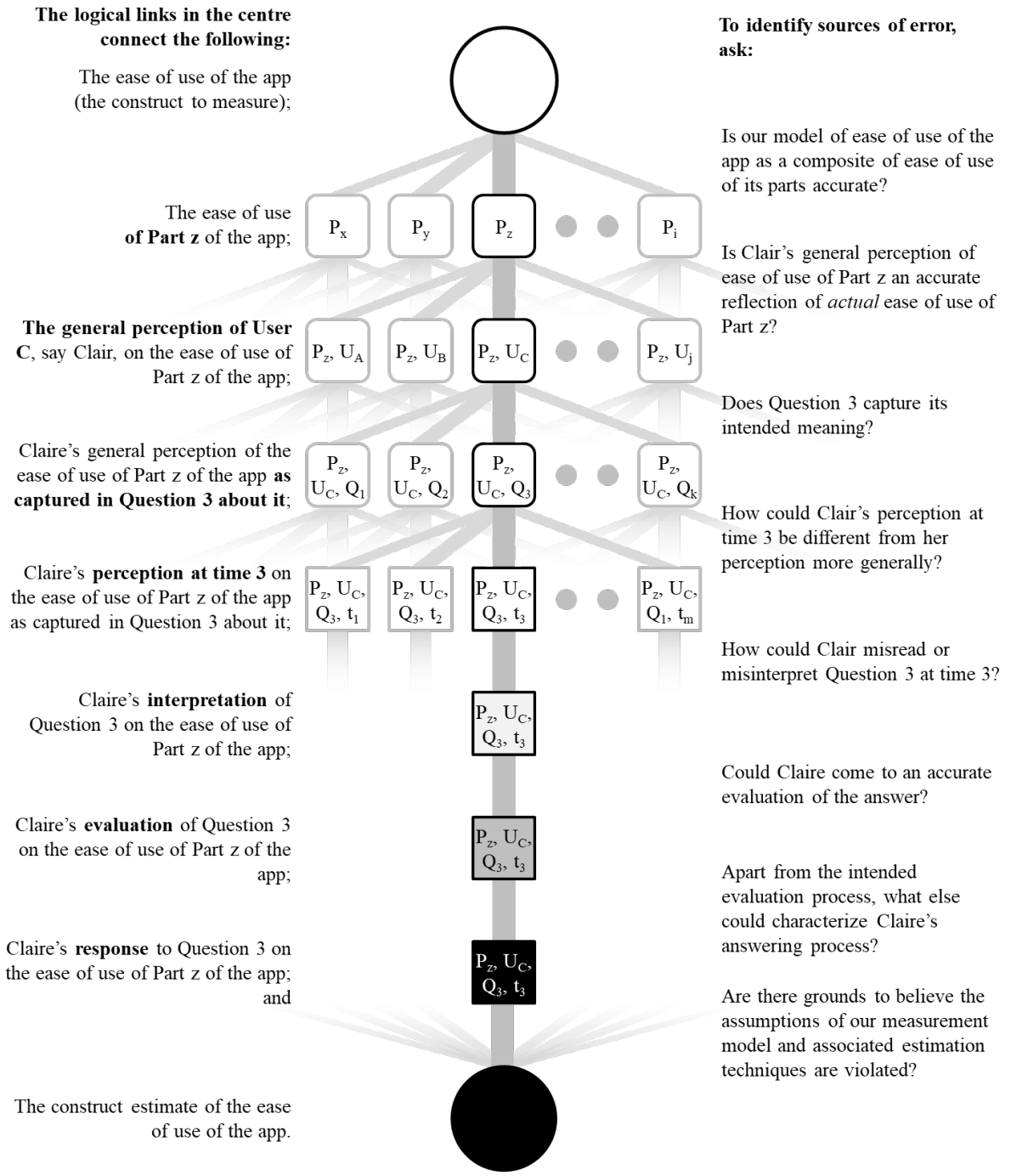

Figure 2: This example illustrates how decomposing the logic of a measurement into consecutive logical links can help researchers identify specific sources of error.

3.3 Tying Logical Links to Sources of Error

Extant knowledge of patterns observed before can inform which sources of error are associated with these links. Table 1 illustrates this by classing sources of error of a measurement according to its consecutive logical links involving one reflective model and one-off self-reported questionnaire data. Table 2 classes common sources of error associated with a variety of logical links that may be applied in a modular fashion in a variety of measurements.

Logical Link	Associated Sources of Error
Construct to Manifestation	A well-defined construct may be inappropriately modelled in terms of its manifestations (Aguirre-Urreta & Marakas, 2012; Jarvis, MacKenzie, & Podsakoff, 2012; Petter, Rai, & Straub, 2012). A proclaimed manifestation may in fact not be consistent with the definition of the construct. It may be caused by other constructs, or the construct may only give rise to it under conditions that are not met, or it may not stand in a relationships to it as modelled (Cook & Campbell, 1979; Rigdon, 2014a).
Manifestation to Question	A question may inappropriately capture a manifestation of a construct. For example, a question may refer to something else or may be unclear (e.g. Dillman, 2000; Netemeyer, Bearden, & Sharma, 2003; Tourangeau et al., 2000). Its response scale may be confusing, inconsistent with the question, or unable to capture accurate answers (e.g. Verhagen, van Den Hooff, & Meents, 2015).
Question to Evaluation	Participants may inappropriately evaluate a question. For example, they may lack the motivation, energy, vocabulary, and other cognitive abilities to do so (Churchill, 1979; Nunnally & Bernstein, 1994; Tourangeau et al., 2000; Viswanathan, 2005). A question may be too difficult, inaudible, or illegible. The instructions may be unclear and the time pressure and incentives may be inappropriate. The time of the day, the location, and the order of the questions may have an unintentional influence on their evaluations (Dillman, 2000; Drury & Farhoomand, 1997; Harrison, McLaughlin, & Coalter, 1996; Podsakoff et al., 2012; Schwarz & Sudman, 1992). A participant may have distracting thoughts and feelings while evaluating a question, due to idiosyncratic associations with certain words, or perceptions of fatigue, hunger, pain, noise, a phone ringing, or even due to simultaneous actions (e.g. Edwards, 2008).
Evaluation to Response	An evaluation may not be reported. Questions may be too sensitive to answer honestly (e.g. Dillman, 2000; Netemeyer et al., 2003; Tourangeau et al., 2000). The participant's anonymity or the lack thereof may affect the honesty of the response. A participant may have certain response tendencies (e.g. Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), or lack the motivation or incentives to provide accurate answers (e.g. Aronson, Wilson, & Brewer, 1998; Podsakoff et al., 2003; Podsakoff et al., 2012; Richman, Kiesler, Weisband, & Drasgow, 1999).
Response to Record	An accurate response may be inappropriately recorded due to its illegibility, a data entry mistake, a technical failure, and so on.
Record to Factor Score	A well-recorded response may not be used appropriately for calculating factor scores as its assumptions may be violated. The specification of the factor model may deviate from the conceptual model, i.e. it may deviate from the specified relations between a construct and the referents of its indicators (e.g. Petter, Straub, & Rai, 2007; Rigdon, 2013).
Factor Score to Estimate	A factor score may be an inappropriate estimate when the assumptions underlying the factor analysis are violated. These typically include local independence, linearity of relationships, and homogeneity of relationships across entities (e.g. Becker, Rai, Ringle, & Völckner, 2013; Havlicek & Peterson, 1977; Jarque & Bera, 1987; Meredith, 1993; Petter et al., 2007). Further, factors scores are indeterminate: their validity depends on the arbitrary method chosen to calculate them (Mulaik, 2010; Rigdon, 2012).

Table 1: Sources of error associated with common logical links in a typical case of measurement

Link	Associated Sources of Error
Construct to Referent	A construct may be inappropriately modelled in terms of its manifestations, effects, constituent parts, dimensions, or other referents of indicators that stand in some relation to the construct (Goertz, 2006; Petter et al., 2012). For example, a specification of its parts may be incomplete, superfluous, or it may contain overlapping parts (Haynes et al., 1995; MacKenzie et al., 2011). A dimensional model may inappropriately specify how dimensions combine to make up the construct (Law et al., 1998).
Referent to Referent	The referent of an indicator may be modelled in terms of its own referents, yielding a hierarchical model (e.g. Edwards, 2001; Law et al., 1998; Polites et al., 2012). These links can suffer from the same sources of errors as the links between the construct to be measured and the referents of its immediate indicators.
Referent to Detection	A referent of an indicator may be implemented through autonomous systems and physical detectors, such as those that automatically log events or aim to detect heart rate, skin conductance, eye movements, gamma waves, and so forth. Errors may stem from the processes of designing, installing, calibrating, and operating the instrument, depending on the specific apparatus. Logs of device use, for example, may be inappropriately ascribed to a principal user and ignore the use of replacements and other alternatives. Further, using detectors to infer referents of more abstract psychological constructs can be problematic (Dimoka, 2012; Fazio & Olson, 2003). For example, we still know little about how to best infer people's stress, affect, and reward from detections of skin conductance, heart rate variability, and activation of the nucleus accumbens (basal forebrain) respectively (e.g. Carlson, 2013).
Referent to Record	A referent of an indicator may be linked directly to a record when relying on past measurements, or 'secondary data', stored in databases, documents, and logs. The sources of error corresponding to this link include all inconsistencies between the original measurement operations and what the result of these operations – i.e. the record – is taken to mean (Burton-Jones & Lee, 2017; Ketchen, Ireland, & Baker, 2013; Wennberg, 2005).
Referent to Stimulus	A referent of an indicator may be inappropriately implemented into a question, picture, sound, or any other linguistic or non-linguistic stimulus. For example, a sound may be inaudible, a question illegible, or the membership of a stimulus to an intended category may be ambiguous (e.g. Dillman, 2000; Greenwald, McGhee, & Schwartz, 1998).
Stimulus to Response	A good stimulus may not produce the appropriate response, when, for example, the instructions are unclear or evoke an inappropriate degree of time pressure, social pressure, or other forms of stress. The lag with which stimuli are presented may obscure the interpretation of response times (e.g. Greenwald et al., 1998).
Record to Estimate	Records may be inappropriately combined to produce estimates. Data cleaning steps may involve clerical errors. Missing data may be inputted inaccurately due to the shortcomings of the procedures (Little & Rubin, 2014; Schafer & Graham, 2002). The mathematical procedures of estimation may involve assumptions that do not hold (Jarvis et al., 2012; Rigdon, 2012; Viswanathan, 2005).
Record to Model Estimate	Records can be used to test the fit of models or the support for inter-construct hypotheses without separately obtaining estimates for constructs. Sources of error are violations to the assumptions underlying these estimation techniques, related to linearity, normality, measurement invariance, or independence of distributions (e.g. Becker et al., 2013; Havlicek & Peterson, 1977; Jarque & Bera, 1987; Meredith, 1993; Mulaik, 2010; Petter et al., 2007; Rigdon, 2012).

Table 2: Sources of Error Associated with Logical Links of Measurement

4 The Management of Sources of Error

Evaluating a plan of measurement and informing its revisions require not only an understanding of where error comes from, but also what the impact of its sources are. This impact depends on how much interference each source introduces at each logical link, and how much of this interference is transformed, e.g. by way of dilution, compensation, or

statistical control. In other words, sources of error are best managed with an understanding of the patterns in where and how much they impact a measurement system.

4.1 Patterns in Error

Sources of error require most attention when they are systemic rather than incidental, i.e. when they affect the entire measurement system, or a subsystem.[7] Figure 1 illustrates examples of sources of error that are incidental (labelled OE2), systematic (labelled CE1, CE2, OE3, IE1), and in between incidental and systematic (labelled OE1). Sources of error are more systemic when they apply to more dimensions of the indicator data, whether they do so at the conceptual, operational, or inferential level. For example, if the indicators refer to effects of a construct, two systemic and conceptual sources of error are (1) the existence of alternative explanations that can account for the effects and (2) the construct failing to cause the effects. Examples of systemic operational sources of error are violated assumptions underlying the indicators, e.g. when respondents do not have the required knowledge, ability, or motive to answer questions accurately. An incidental operational error may be due to one question being ambiguous. An example of a systemic source of error at the inferential level is a violation to the assumptions underlying estimation, causing new interference across estimates.

Patterns of systemic error can be related to not only the aspects of a chosen measurement method (e.g. particular items, instruments, assessments or estimation procedures), also with the meaning of the construct itself. Figure 3 illustrates this with a joint distribution of four separate continuous indicators and the continuous construct they refer to. Here, error in Indicator A is independent of the construct, while error in Indicators B, C, and D co-varies with the construct.

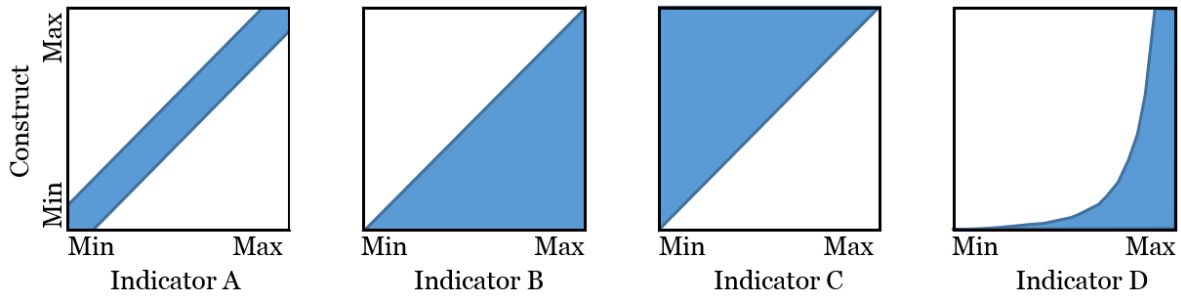

Figure 3: Illustrations of joint distributions of indicators and the construct they refer to (both continuous). Error in Indicator A is uniform, while error in Indicators B, C, and D co-vary with the construct.

Consider one specific source of error in a measurement where an indicator refers to the frequency of an event that is an effect of the construct, and the effect could also be triggered by other causes. Here, the joint distribution might be similar to that of Indicator B in Figure 3, where high levels of the indicator tell us little about the construct. If the indicator would feature in a standard factor analysis, this error would violate the assumption that

[7] We can speak of subsystems when constructs are modelled as consisting of multiple sub-constructs, when data is collected repeatedly, or from multiple informants, or when data is multidimensional for another reasons.

measurement error is normally and independently distributed as the level of error depends on the level of the construct.

Another example is the bias introduced through socially desirability. When the response scale of an indicator refers to behaviours that vary in their social desirability, the values corresponding to the least socially desirable should be more indicative of the actual behaviours than those corresponding to the others. Similarly, when, for example, a computing skill is being tested with a quiz, and a question is very easy, only wrong answers are indicative of the level of skill. This may follow a joint distribution similar to Indicator D in Figure 3.

In the case of categorical variables, such as a condition being either present or absent, there too can patterns in error co-vary with true values. For example, an indicator will have high sensitivity (i.e. high true positive rate) and low specificity (i.e. low true negative rate), when it is based on manifestations that are necessary but not sufficient. Vice versa, an indicator will have low sensitivity and high specificity when it relies on a manifestation that is sufficient but not necessary.

Anticipating these patterns in potential measurement error can help inform the remedies that can best improve the validity of a plan of measurement.

4.2 Remedies

Whether a source of error is associated with one indicator or an entire system of measurement, four categories of remedies may be considered: modify, control, add, and drop.

4.2.1 Modify

The most preferred remedy is to modify a plan of measurement such that it no longer is sensitive to the source of error, without exposing it to another. For example, Indicator B, C, or D from Figure 3 would be improved if it could be modified such that it would behave according to Indicator A. For example, when the referent of a categorical indicator is necessary but not sufficient to indicate a particular construct, modifying it to one that is both necessary and sufficient would improve its usefulness. To illustrate, consider the item "Based on my experience with the online vendor in the past, I know it is trustworthy" used to indicate trust in the online vendor (Gefen et al., 2003). Arguably, one might trust online vendors based on reputation, reviews, or certificates, rather than one's experience with it, and thus the indicator may underestimate trust. Modifying a set of items by including a complete range of trust sources, or removing the references to these sources altogether, will remove this particular source of error.

While various guidelines in the literature help researchers deal with specific item-specific issues (Clark & Watson, 1995; Dillman, 2000; Haynes et al., 1995; MacKenzie et al., 2011; Tourangeau et al., 2000), protecting them against interference is easier said than done (Spector, 2006). Interference tends to stem from a complex and hidden interplay of contextual, idiosyncratic, and circumstantial factors. For example, a researcher may not know fully the conditions under which a construct gives rise to its manifestations. It may be impossible to ensure that participants interpret all questions as intended and report answers accurately. Sometimes, shielding the measurement process against one threat exposes it to another. For example, indirect questioning may help prevent socially desirable responses but it may capture content outside of the construct's domain (Fisher, 1993). One may wish to rely on methods different from question-and-answer by logging events or recording response times, eye movements, skin conductance, or other physical phenomena (Bradley, Greenwald, Petry,

& Lang, 1992; Carlson, 2013; Segerstrom & Nes, 2007). As shown in Table 2, such alternative means may suffer from their own sources of error that are hard to prevent. Inevitably, potential threats to validity can be found along the entire logical relationship between a construct and its estimate. In Spector's (2006, p230) words, "each operationalization of a variable or method-trait combination carries with it a unique set of potential biases."

4.2.2 Control

A second category of remedies is to deal with the interference from a given source of error and control for it, through measured or unmeasured approaches. For example, instruments that measure a respondent's sensitivity to give social desirable rather than honest answers can be used as a measured control for social desirability bias (Paulhus, 1988). Similarly, other response tendencies such as acquiescence can be measured to allow for statistical control (Paulhus, 1991; Winkler, Kanouse, & Ware, 1982). While such measured approaches are more precise than unmeasured ones, they do impose an operational burden; it is impractical to measure many sources of error.

Unmeasured approaches to control for error are common in factor analytic approaches, where error is modelled as affecting the indicator data. Typically, in factor analysis each unique indicator, such as a questionnaire item, is modelled as affected by random error that is completely independent and normally distributed around zero (Harman, 1976; Nunnally & Bernstein, 1994). While this does account for sources of error that behave accordingly, in many measurements indicators resemble one another – they may rely on the same questionnaire, assessment, response scale, and relation to the construct – often meaning that they share sources of error. Sometimes, such common method error is modelled as affecting a set of indicators (Harman, 1976; Lindell & Whitney, 2001; Tehseen, Ramayah, & Sajilan, 2017; Williams, Hartman, & Cavazotte, 2010). While popular, this technique can be problematic, as it relies on various assumptions (Chin, Thatcher, & Wright, 2012). A common assumption is that the common method error can be modelled as a unitary construct, which limits the ability to control for separate common sources of error, such as both acquiescence and social desirability. While unmeasured approaches to control for sources of error do not directly impact the operations of measurement, the more realistic their models, the more data is needed to identify these models statistically.

Hence, measured and unmeasured approaches come with operational costs and constraints respectively. Knowledge of the specific sources of error and their behaviour will help inform how to implement these approaches with reasonable assumptions, to the best effect.

4.2.3 Add

Another way to mitigate the impact of sources of error is by introducing indicators, methods, or other measurement subsystems, that are less sensitive to these sources of error (Campbell, 1957; Podsakoff et al., 2012). As a result, these sources of error become less systemic and incidental to a smaller part of the measurement plan; error would be diluted (Peter, 1981).

Ways to diversify a set of indicators include appending self-report with peer-report; a single survey with the momentary assessment method; and a measurement of a construct as a sum of its parts with an indicator that refers to the construct as a cause of its effects. Diversifying indicators works well when new indicators are not clearly inferior to extant ones (Burton-Jones, 2009), but complement them in terms of their 'error profile'. The fewer indicators a source of error affects, the more limited its impact on the validity of measurement (Burton-

Jones, 2009; Houts, Cook, & Shadish, 1986; Nunnally & Bernstein, 1994). As a peculiarity of just one indicator a source of error often has the least impact on the validity of measurement, depending on the reliance on that indicator in the method of estimation.

Apart from diversifying a set of indicators, one can also specifically target a new indicator that is less sensitive to a given source of error. For example, Indicators B and C in Figure 3 complement one another by being sensitive to opposite ends of the scale of the construct. Adding C to B or conversely may be especially worthwhile if either cannot be changed in an indicator that behaves more like Indicator A, noting that this non-traditional use of indicators does require compatible estimation techniques. Consider, for example, two items in a computer aptitude test, one difficult – and thus indicative of aptitude only at the higher end – and one easy, and thus indicative only at the lower end. Such items can be combined with models based on Item Response Theory such that the combined test can be indicative of aptitude along its entire scale (Embretson & Reise, 2013).

While introducing dissimilar indicators may be operationally expensive, it can effectively reduce the impact of multiple sources of error that are specific to the extant set of indicators. If the new indicators themselves do not introduce a problematic degree of error, this approach is especially worthwhile.

4.2.4 Drop

In some cases, simply dropping an indicator is an effective way of removing error (MacKenzie et al., 2011). It can only be done, however, when the indicator is redundant in terms of its referent and method. That is, it cannot be culled if either its meaning or the means through which it accesses this meaning plays a critical role in the logic of measurement.

In sum, the degree with which sources of error undermine the validity of measurement depends on where they interfere, how severely they interfere, and how much of their impact can be reduced. Often, their presence is inevitable and their impact elusive. Yet even while deficient, a thorough evaluation of these sources of error should inform the design of measurement as it can lead to better validity.

5 Discussion

This paper has provided a framework for unpacking the logical relationship of measurement, used that framework to evaluate and classify sources of error, and provided a strategy for the identification of remedies based on this classification. We believe the framework, classification, and the strategy can guide researchers as it builds on and synthesizes existing literature.

By decomposing the logic of measurement into conceptual, operational, and inferential relationships, the framework ties in literature that is often unjustifiably disjointed. Work on constructs (e.g. Barki, 2008; Goertz, 2006) and conceptual modelling (e.g. Polites et al., 2012), including the discussion on reflective and formative models (e.g. Jarvis et al., 2012), is often disjointed from literature on the practical operations of measurement, even though it carries direct implications for the operations of measurement (e.g. "what questions can we ask and are these any good?"). Similarly, while the literature on estimation and the use of statistical techniques and tests is reliant on broad assumptions regarding the conceptual and operational relationships, it rarely examines these (Rigdon, 2014b; Zyphur & Pierides, 2017). Whether they are conceptual, operational, or inferential, sources of error are manifold and entangled; dealing

with one often raises other concerns. Appropriate management of these thus requires a holistic, systemic view.

Our framework, classification, and strategy help with this in various ways. First, decomposing the logical relationship of measurement into conceptual, operational, and inferential relationships, the framework eases a complete and systematic evaluation of sources of error, since the relationships are mutually exclusive and completely exhaustive. As demonstrated with the classification of error, this evaluation can be further eased by breaking down these relationships into elementary logical links, allowing for a measurement-specific classification of potential error. Tying the identification of sources of error to this logical chain of measurement is arguably more intuitive than the more traditional approach of tying it to entities within measurement such as constructs or indicators – see e.g. the discussion on construct-level and item-level error in MacKenzie et al. (2011). Ultimately, sources of error affect the processes of measurement or the relationships between such entities, not entities themselves. Therefore, our recommendations should further help researchers in evaluating sources of error beyond the help provided by existing error classification frameworks.

Second, the recommendations we offer stimulate a complete evaluation of sources of error rather than a more stepwise or ad-hoc evaluation. For example, past guidance in the development of one-off questionnaire instruments has recommended the generation of items first, then the specification of the measurement model, and later still the estimation procedures, with the consideration of potential sources of error at each step (MacKenzie et al., 2011). While these procedures do allow for iteration and subsequent refinement, at each step only narrow, local remedies are sought, under the implicit assumption the measurement approach has been set in stone. Common assumptions are that a single assessment is to take place, to have a single informant per instantiation, to use a questionnaire, to use reflective models, and to model the indicators as linear combinations of the construct and normally distributed error (Churchill, 1979; DeVellis, 2003; MacKenzie et al., 2011). These choices are limiting themselves (Burton-Jones, 2009; Podsakoff et al., 2003; Rigdon, 2013), and assuming them implicitly does not inspire a holistic view on the entire measurement plan. Our recommendations encourage taking such a view, and revising measurement plans through its models, data collection techniques, informants, assessments, estimation techniques etc. In particular, they encourage adding methods that are maximally different, as these are least likely to share sources of error (Campbell & Fiske, 1959; Peter, 1981), and best help to triangulate measurement. At the very least, being method-agnostic, the recommendations may alleviate the issue of dogmatic application of ready-made formulas (Zyphur & Pierides, 2017).

Relatedly, our recommendations help advance the consideration of measurement error in the development process. Most of the discussion of measurement error in the literature assumes data has been collected, when considerable effort has been spent and opportunities for avoiding and alleviating sources of error have narrowed. It focuses on statistical measures of various forms of validity, or lack thereof, and not on taking measures to improve the actual validity of measurement (Zyphur & Pierides, 2017). While less quantitative, a-priori evaluation of potential error can thus be more instructive.

This paper has taken a first step in integrating extant insights into sources of error and offering recommendations to researchers with measurement plans for evaluating a-priori and systematically potential sources of error, whether these are conceptual, operational, or inferential. Such a comprehensive evaluation provides a basis for improving a plan of

measurement in an integrated, rather than ad-hoc manner, and allow for the identification of remedies that address multiple sources of error at once. As such, this strategy can complement and enhance existing approaches toward the development of measurement (MacKenzie et al., 2011).

Future steps to better measurement guidance could help researchers exploit practical opportunity for creative measurement solutions and simultaneously follow well-established best practices (Burton-Jones, 2009). For example, much measurement in the academic behavioural disciplines is part of a broader attempt to confirm structural models, where structural equation modelling is a well-established approach to do so. While it is often implemented with reflective measurement models where each indicator corresponds to one question on a manifestation of its construct, asked in a one-off self-report questionnaire, there are many possible variations. Within that approach, some literature has dealt with the use of alternative measurement models (Diamantopoulos, 2011; Diamantopoulos & Temme, 2013; Polites et al., 2012), and other literature has dealt with the use of indicator data from repeated assessment (Crowder, 2017; Muthén, 2002). A synthesis of streams of work like these positioned in the context of measurement development could help researchers identify and evaluate a broad range of possibilities, such that measurement can suit its circumstances (Zyphur & Pierides, 2017). This would more clearly show possible responses to the evaluation of sources of error that undermine a more traditional plan of measurement.

Hence, our hope is that this paper will inspide further steps in the development of guidelines that help researchers anticipate, avoid, and alleviate source of measurement error. It should pave the way for well-designed, valid measurement within our disciplines.

References

Aguirre-Urreta, M. I., & Marakas, G. M. (2012). Revisiting Bias Due to Construct Misspecification: Different Results from Considering Coefficients in Standardized Form. *MIS quarterly, 36*(1), 123-138.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology.

Bagozzi, R. P. (2011). Measurement and Meaning in Information Systems and Organizational Research: Methodological and Philosophical Foundations. *MIS quarterly, 35*(2), 261-292.

Barki, H. (2008). Thar's gold in them thar constructs. *ACM SIGMIS Database, 39*(3), 9-20.

Becker, J.-M., Rai, A., Ringle, C. M., & Völckner, F. (2013). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS quarterly, 37*(3), 665-694.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305.

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition, 18*(2), 379.

Burton-Jones, A. (2009). Minimizing method bias through programmatic research. *MIS quarterly, 33*(3), 445-471.

Burton-Jones, A., & Lee, A. S. (2017). Thinking About Measures and Measurement in Positivist Research: A Proposal for Refocusing on Fundamentals. *Information systems research, 28*(3), 451-467.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Carlson, N. (2013). *Physiology of Behavior* (11 ed.). New Jersey: Pearson Education, Inc.

Chin, W. W., Thatcher, J. B., & Wright, R. T. (2012). Assessing common method bias: problems with the ULMC technique. *MIS quarterly, 36*(3), 1003-1019.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research, 16*(1), 64-73.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*: Rand McNally.

Cote, J. A., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research, 14*(4), 579-582.

Cronbach, L. J. (1989). Construct validation after thirty years. *Intelligence: Measurement, theory, and public policy, 3*, 147-171.

Crowder, M. (2017). *Analysis of repeated measures*: Routledge.

DeVellis, R. F. (1991). *Scale Development: Theory and Applications*. Newbury Park, CA: Sage Publications.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (Vol. 2nd). Thousand Oaks, California: Sage Publications.

Diamantopoulos, A. (2011). Incorporating Formative Measures into Covariance-Based Structural Equation Models. *MIS quarterly, 35*(2), 335-358.

Diamantopoulos, A., & Temme, D. (2013). MIMIC models, formative indicators and the joys of research. *AMS review, 3*(3), 160-170.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (Vol. 2): Wiley New York.

Dimoka, A. (2012). How to conduct a functional magnetic resonance (fMRI) study in social science research. *MIS quarterly, 36*(3), 811-840.

Drury, D. H., & Farhoomand, A. (1997). Improving management information systems research: question order effects in surveys. *Information Systems Journal, 7*(3), 241-251.

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational research methods, 4*(2), 144-192.

Edwards, J. R. (2008). To prosper, organizational psychology should… overcome methodological barriers to progress. *Journal of Organizational Behavior, 29*(4), 469-491.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological methods, 5*(2), 155.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*: Psychology Press.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual review of psychology, 54*(1), 297-327.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 303-315.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. *MIS quarterly, 27*(1), 51-90.

Goertz, G. (2006). *Social science concepts: A user's guide*: Princeton University Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464.

Gregor, S., & Klein, G. (2014). Eight obstacles to overcome in the theory testing genre. *Journal of the Association for Information Systems, 15*(11), I.

Harman, H. H. (1976). *Modern factor analysis* (Vol. 3rd). Chicago: University of Chicago Press.

Harrison, D. A., McLaughlin, M. E., & Coalter, T. M. (1996). Context, cognition, and common method variance: Psychometric and verbal protocol evidence. *Organizational behavior and human decision processes, 68*(3), 246-261.

Havlicek, L. L., & Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the Pearson r. *Psychological Bulletin, 84*(2), 373.

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

Houts, A. C., Cook, T. D., & Shadish, W. R. (1986). The person-situation debate: A critical multiplist perspective. *Journal of personality, 54*(1), 52-105.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique, 55*(2), 163-172.

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

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2012). The Negative Consequences of Measurement Model Misspecification: A Response to Aguirre-Urreta and Marakas. *MIS quarterly, 36*(1), 139-146.

Ketchen, D. J., Ireland, R. D., & Baker, L. T. (2013). The Use of Archival Proxies in Strategic Management Studies Castles Made of Sand? *Organizational research methods, 16*(1), 32-42.

Law, K. S., Wong, C.-S., & Mobley, W. M. (1998). Toward a taxonomy of multidimensional constructs. *Academy of Management Review, 23*(4), 741-755.

Lewis, B. R., Templeton, G. F., & Byrd, T. A. (2005). A methodology for construct development in MIS research. *European Journal of Information Systems, 14*(4), 388-400.

Lindell, M. K., & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology, 86*(1), 114-121.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333): John Wiley & Sons.

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

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*: Routledge.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*(2), 300.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Mulaik, S. A. (2010). *Foundations of Factor Analysis* (2nd edition ed.). Boca Raton, FL: Taylor and Francis Group.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3 ed.). New York, NY: McGraw-Hill.

O'Leary-Kelly, S. W., & Vokurka, R. J. (1998). The empirical assessment of construct validity. *Journal of operations management, 16*(4), 387-405.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657-660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716-4711-4718.

Paulhus, D. (1988). Balanced inventory of desirable responding (BIDR). *Acceptance and Commitment Therapy. Measures Package, 41*.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1. Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA, US: Academic Press.

Peter, J. P. (1981). Construct validity: A review of basic issues and marketing practices. *Journal of marketing research*, 133-145.

Petter, S., Rai, A., & Straub, D. (2012). The critical importance of construct measurement specification: a response to Aguirre-Urreta and Marakas. *MIS quarterly, 36*(1), 147-155.

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

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & 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), 879-903.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual review of psychology, 63*, 539-569.

Polites, G. L., Roberts, N., & Thatcher, J. (2012). Conceptualizing models using multidimensional constructs: a review and guidelines for their use. *European Journal of Information Systems, 21*(1), 22-48.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754.

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: in praise of simple methods. *Long Range Planning, 45*(5), 341-358.

Rigdon, E. E. (2013). Lee, Cadogan, and Chamberlain: an excellent point... But what about that iceberg? *AMS review, 3*(1), 24-29.

Rigdon, E. E. (2014a). Comment on "Improper use of endogenous formative variables". *Journal of Business Research, 67*(1), 2800-2802.

Rigdon, E. E. (2014b). Rethinking partial least squares path modeling: breaking chains and forging ahead. *Long Range Planning, 47*(3), 161-167.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147.

Schwarz, N., & Sudman, S. (1992). *Context Effects in Social and Psychological Research*. New York: Springer-Verlag.

Segerstrom, S. C., & Nes, L. S. (2007). Heart rate variability reflects self-regulatory strength, effort, and fatigue. *Psychological Science, 18*(3), 275-281.

Servick, K. (2018, 27 August 2018). 'Generous' approach to replication confirms many high-profile social science findings. *Science*.

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), 221-232.

Tehseen, S., Ramayah, T., & Sajilan, S. (2017). Testing and controlling for common method variance: a review of available methods. *Journal of Management Sciences, 4*(2), 142-168.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*: Cambridge University Press.

Verhagen, T., van Den Hooff, B., & Meents, S. (2015). Toward a Better Use of the Semantic Differential in IS Research: An Integrative Framework of Suggested Action. *Journal of the Association for Information Systems, 16*(2), 108-143.

Viswanathan, M. (2005). *Measurement error and research design*: Sage.

Wennberg, K. (2005). Entrepreneurship research through databases: Measurement and design issues. *New England Journal of Entrepreneurship, 8*(2), 9-19.

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

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67*(5), 555.

Zwanenburg, S. P. (2015). *How to Tie a Construct to Indicators: Guidelines for Valid Measurement.* Paper presented at the International Conference on Information Systems, Fort Worth.

Zyphur, M. J., & Pierides, D. C. (2017). Is quantitative research ethical? Tools for ethically practicing, evaluating, and using quantitative research. *Journal of Business Ethics, 143*(1), 1-16.