Using single-item measures for construct measurement in management research

Conceptual issues and application guidelines

Schlüsselbegriffe
Messetheorie; Reliabilität; Single-Item Skalen; Validität

Keywords
Measurement; reliability; single-item measures; validity

Zusammenfassung

Abstract
The use of single-item measures in management research has been subject to heavy debate in recent literature. This paper provides researchers with concrete guidelines on how to assess the extent to which a single-item measure can be legitimately used to operationalize the focal construct. We first present a conceptual perspective on single-item measures and follow this by an integrative framework within which the potential acceptability of single-item measures can be comprehensively evaluated on multiple relevant criteria.

Autoren
Christoph Fuchs, Assistant Professor, Department of Marketing and Statistics, Aarhus School of Business, University of Aarhus, Haslehøjsgårdsvej 10, DK-8210 Aarhus, Denmark, Phone: +45-8948/6494, Fax: +45-8615/3988, E-Mail: chfu@asb.dk; Adamantios Diamantopoulos, Professor of International Marketing, Department of International Marketing, University of Vienna, Brunnner Strasse 72, A-1210 Vienna, Austria, Phone: +43-01/4277-38031, Fax: +43-1/4277-38034, E-Mail: adamantios.diamantopoulos@univie.ac.at.
1. Introduction

Surveys are the primary data collection method in management research (Bryman and Bell, 2003). The typical survey in this field seeks to generate data for testing relationships among several constructs; thus, a single construct is rarely covered in isolation. In this context, besides constructs being conceptualized as dependent and independent variables, most research designs also contain control and/or moderator variables, each of them usually measured with multi-item scales (e.g., Krosnick, 1999). This requires very lengthy survey instruments, which are likely to overload respondents (e.g., Wanous, Reichers and Hudy, 1997), lead to a decrease in response rates, break-offs, and contain more missing values (e.g., Dillman, Sinclair, and Clark, 1993). Furthermore, long questionnaires tend to lose respondents who are not very involved or interested in the investigated domain, which eventually may lead to a sampling bias (Moore et al., 2002). From a practical perspective, long measurement instruments are more costly because additional expenses for data collection and coding need to be covered (Moore et al., 2002). Overall, since space on a questionnaire is limited and therefore very valuable, researchers frequently face pressure when deciding which constructs to include or to exclude in a study; all other things being equal, the more constructs that are included, the greater the demands on measurement in terms of number of items to be included in the survey questionnaire.

Rather than omit potentially relevant constructs (and hence risk misspecification of theoretical models), researchers may opt for the use of shortened versions of existing multi-item scales or even employ single-item measures to operationalize (some of) their constructs. Indeed, several researchers in psychology have highlighted the need of making measurement more efficient by either substantially reducing the length of the scales applied or by using single-item measures as opposed to multiple item measures (e.g., Stanton et al., 2001; Nagy, 2002; Russell et al, 2004). With specific reference to the latter, single-item measures offer the important advantages of being short, flexible and easy to administer (Pomeroy, Clark and Philip, 2001). They are also less time consuming and not monotonous to complete (Gardner et al., 1998), thus reducing response biases (Drolet and Morrison, 2001). However, conventional measurement wisdom in business and management research strongly advocates the use of multi-item scales (e.g., see Boyd, Gove and Hyatt, 2005) and, indeed, practically all measure development textbooks (e.g., DeVellis, 2003; Netemeyer, Bearden and Sharma, 2003; Viswanathan, 2005) and articles (e.g., Churchill, 1979; Comrey, 1988; Homburg and Giering, 1996; Diamantopoulos and Winklhofer, 2001) focus on the construction of multi-item measures. Accordingly, editors and reviewers of academic journals are reluctant to accept manuscripts using single-item measures to operationalize some of their constructs (Singh, 2003); in fact, »the use of single-item measures in academic research is often considered a fatal flaw in the review process« (Wanous, Reichers and Hudy, 1997, p. 247).

Recently, however, several authors in the management field have challenged such conventional wisdom by demonstrating that single-item measures can have acceptable psychometric properties and are, therefore, a potentially viable alternative to multi-item scales for construct measurement purposes (e.g., Drolet and Morrison, 2001; Bergkvist and Rossiter, 2007). Consequently, »the use of single-item measures should not be considered fatal flaws in the review process. Rather, their appropriateness for a particular piece of research should be evaluated« (Wanous, Reichers and Hudy, 1997, pp. 250–251, added emphasis). Unfortunately, there is currently little guidance to practicing researchers in the management field regarding the circumstances under which the use of single-item measures may be acceptable. This is both because most methodological work on single-item measures has been conducted outside the management discipline and because different authors have focused on specific aspects of single-item measures (e.g., reliability or predictive validity) in isolation. What is missing is an integrative framework within which the potential acceptability of single-item measures can be comprehensively evaluated on multiple relevant criteria.

It is this gap in the literature that the present paper seeks to fill. Specifically, drawing from an interdisciplinary literature review on the use of single-item measures, we aim to identify the con-
quisitions under which their use can be justified in substantive empirical research. Our objective is to provide management researchers with concrete guidelines on how to assess the extent to which a single-item measure can be legitimately used to operationalize the focal construct.

In the section that follows, we provide a conceptual perspective on single-item measures and follow this with a discussion of reliability and validity assessment issues. Next, we present key considerations influencing the choice of single- vs. multi-item measures and conclude the paper with some suggestions regarding future research on single-item measures. Throughout the following discussion, a single-item measure is considered to be an individual measure or indicator as defined under the »total disaggregation« model of construct representation by Bagozzi and Heatherton (1994). In other words, a single-item measure, as conceived here, cannot be further decomposed to lower-level constituents as is the case, for example, with composite indicators made up of several items which are aggregated/averaged to compute the composite. An illustrative example of a single-item measures is »All in all, would you say that you are satisfied or dissatisfied with your job?«, scored on, say, a 7-point scale ranging from »very dissatisfied« to »very satisfied«.

2. A Conceptual Perspective on Single-Item Measures

2.1. Reflective Measurement

Given a focal construct $\eta$ and a measure of it $x_1$, a single-item operationalization under a reflective measurement perspective can be described as follows:

$$ x_1 = \lambda_1 \eta + \epsilon_1 $$

(1)

where $\lambda_1$ is the loading of $x_1$ on $\eta$ and $\epsilon_1$ is the measurement error associated with $x_1$. It is assumed that $\text{COV}(\eta, \epsilon_1) = 0$ and that $\text{E}(\epsilon_1) = 0$. Note that equation (1) assumes that $x_1$ is a unidimensional item, that is, only a single latent variable (i.e., $\eta$) contributes to its variation.2

There are two ways one can conceptually approach equation (1). The first way assumes that $x_1$ is the only measure that can be used to represent $\eta$ or, what amounts to the same thing, that changes in $\eta$ are solely and exclusively reflected in changes in $x_1$. Here one faces the problem that it is not possible to estimate the parameters $\lambda_1$ and $\theta_1$ (the variance of $\epsilon_1$) because the model in equation (1) is underidentified.3 The usual «fix» in this case is to either assume that $x_1$ is a perfect measure of $\eta$ (which implies an error variance of zero, i.e., $\theta_1 = 0$) or to assume a certain level of reliability for $x_1$ and use this to set the error variance to a specific value (Jöreskog and Sörbom, 1989); for example, if a reliability of .80 is assumed, $\theta_1$ would be constrained to $(1-.80) \times \text{VAR}(x_1)$.

Neither of these alternatives is particularly attractive from a conceptual perspective. Assuming complete absence of measurement error implies that »a concept becomes its measure and has no theoretical meaning beyond that measure« (Bagozzi, 1982, p. 15). Assuming a certain amount of measurement error is a better option but still unsatisfactory because there is no way of knowing/testing whether the chosen value is accurate or not. Given that a measure’s reliability (or lack of) impacts its relationships with other measures (see Nunnally and Bernstein, 1994), depending on the reliability values chosen to fix $\theta_1$, conclusions about the substantive linkages between $\eta$ and other constructs may well differ.

A second problem arises because it is difficult to provide a convincing answer to the question why is $x_1$ the only indicator that can be used to operationalize $\eta$, or put differently, why no other indicator(s) can be possibly used. In this context, »an observable measure never fully exhausts everything that is meant by a construct... a measure of a construct could never have complete content validity« (Peter, 1981, p. 134). Why cannot a second indicator, say $x_2$, also capture the part of the meaning of $\eta$ that is already captured by $x_1$ and/or that part which is not captured by $x_1$? In other words, what is it that makes $x_1$ unique? A possible answer to this question is that it is sometimes «difficult to generate

---

2 We focus initially on the unidimensional case, because of the acknowledged desirability of unidimensionality as a measurement property in management research (e.g., see Anderson and Gerbing, 1982; Gerbing and Anderson, 1988; Danes and Mann, 1984; Steenkamp and van Trijp, 1991). We will consider the case of multiple latent variables impacting on $x_1$ later in this section.

3 Assuming that the variance of $\eta$ is standardized, there are two parameters to be estimated, namely $\lambda_1$ and $\theta_1 = \text{VAR}(\epsilon_1)$, but only one piece of information (the variance of $x_1$).
multiple items to measure a construct. The major
reason is that the construct is simple and single-
faceted, and it is impossible to create many differ-
ent items that measure the same underlying con-
struct (Poon, Leung and Lee, 2002, p. 276, added
emphasis). Indeed, within the context of the C-
OAR-SE scale development procedure, Rossiter
(2002, p.331) argues that »a concrete singular ob-
ject to be rated in terms of a concrete attribute
needs only a single-item scale«.

An alternative way one can approach equation
(1) is by considering $x_1$ as one of a set of indicators
that could be potentially used to measure $\eta$. Here,$x_1$ is not seen as the measure of $\eta$ but as a repre-
sentative measure of $\eta$. This viewpoint is much
more consistent with the domain sampling model
in measurement theory which considers »any par-
ticular measure to be composed of responses to a
random sample of items from a hypothetical do-
main of items« (Nunnally and Bernstein, 1994,
p. 216). A single-item measure such as $x_1$ can thus
be considered to be one possible measure of $\eta$ drawn from the relevant domain. Bearing in mind
that »one major source of measurement error oc-
curs when the sampling of the domain items is in-
adequate« (Grapentine, 2001, p.156), one can ask
the question whether a single item can ensure ad-
equate domain representation. This is essentially a
sampling question and according to sampling the-
ory (e.g., see Sudman, 1976; Cochran, 1977), a key
consideration is the extent of variability/heteroge-
neity in the target population; all other things be-
ing equal, the more homogeneous the population,
the smaller the needed sample size. In the extreme,
if all population elements are identical to each
other, then a sample of one is adequate to accu-
rately represent the population.

Applying this logic to the domain sampling
model, $x_1$ could be considered as being representa-
tive of the domain of relevant items if the other
items were similarly related to $\eta$ as $x_1$. This applies
in the case of parallel items, whereby »each item of
a scale is precisely as good a measure of the latent
variable as any other of the scale items... each
item’s relationship to the latent variable is pre-

\[ x_2 - \lambda_2 \eta + \varepsilon_2 \]

(2)

with $\lambda_2$ representing the loading of $x_2$ on $\eta$ and $\varepsilon_2$
measurement error. Given that under the assump-
tion of parallelism $\lambda_1 = \lambda_2$ and $\theta_1 = \theta_2$, only two
parameters need to be estimated (i.e., one loading
and one error variance) from three pieces of infor-
mation (i.e., the variance of $x_1$, the variance of $x_2$,
and the covariance between $x_1$ and $x_2$). Thus the
measurement model defined by equations (1) and
(2) becomes overidentified (with 1 degree of free-
dom) and can be estimated/tested. Subsequently,
either $x_1$ or $x_2$ could be used as a single-item meas-
ure of $\eta$ in a larger model (including antecedents
and/or consequences of $\eta$) with the error variance
of the chosen indicator (i.e., either $\theta_1$ or $\theta_2$) fixed
at the value obtained during the estimation of the
parallel model. Note, in this context, that $x_1$ and $x_2$
(and, for that matter, all other parallel items to $x_1$
and $x_2$) are perfectly interchangeable and empiri-
cally indistinguishable and can only differ in terms
of their specific wording (i.e., their measurement
properties are identical). Of course, one runs the
(very real) danger that $x_1$ and $x_2$ are merely seman-
tically identical (i.e., redundant) items, whereby
»essentially the same item is rephrased in several
different ways« (Boyle, 1991, p. 281). In this case,
the second item in such a pair does not represent
additional sampling from the content domain, and
so one aspect of the domain may be oversampled
(Smith and McCarthy, 1995, p. 306).

One must also pose the question whether the
assumption of parallel items is too restrictive when
operationalizing constructs in management re-
search. The measurement tradition in this research
field has been Jöreskog’s (1971) congeneric model
which merely assumes »that all the items share a
common latent variable. They need not bear equally
strong relationships to the latent variable, and their
error variances need not be equal« (DeVellis, 2003,
p. 25). Thus, unlike with parallel items, under con-
generic measurement assumptions, items are not identical to each other as far as their measurement properties are concerned. This complicates the definition of a »representative« item to be used as a single-item measure of the focal construct. This is a conceptual problem which is not easily resolved even if one has previously developed a psychometrically-sound, multi-item scale following established measurement development procedures (e.g., Churchill, 1979; Spector, 1992; Netemeyer, Bearden and Sharma, 2003). For example, choosing the item with the highest reliability or the item with highest loading (as done, for example, by Loo, 2002) does not conceptually explain why such an item is more representative of the construct domain than the other items (thus making it more suitable as a candidate for a single-item measure). Moreover, what may prove to be the »best« item judged on the above criteria in one sample may well turn out to be not the »best« one in a different sample. Purely relying on empirical results to identify the »best« item is bound to capitalize on the idiosyncrasies of the sample data at hand and, short of replicating the full scale on several samples, there does not appear to be an easy way out of this problem.

Finally, if the unidimensionality assumption is relaxed, then a single-item reflective measurement model becomes particularly problematic. Consider model shown in equation (3). Here, variation in $x_1$ is attributable to two latent variables $\eta_1$ and $\eta_2$, which means that $x_1$ measures two constructs rather than a single construct as was assumed so far. In the absence of any other measures for $\eta_1$ and $\eta_2$, it is not possible to identify which of the latent variables is responsible for the covariation with the other variables in the system. In addition to the obvious identification problems of the model shown in equation (1), there is a major conceptual difficulty with accepting that a single item can be a sufficient measure for multiple constructs.

\[ x_1 = \lambda_1 \eta_1 + \lambda_2 \eta_2 + \varepsilon_1 \quad (3) \]

2.2. Formative Measurement

A single-item operationalization under a formative perspective can be described as follows:

\[ \eta = \gamma_1 x_1 + \zeta \quad (4) \]

where $\gamma_1$ is the expected effect of $x_1$ on $\eta$ and $\zeta$ is a disturbance term. It is assumed that $\text{COV}(x_1, \zeta) = 0$ and $E(\zeta) = 0$.

Unlike with the reflective measurement perspective discussed in the previous section, there is only one interpretation of equation (4), namely that $\eta$ is determined by $x_1$ as well as by $\zeta_i$, the latter capturing the impact of all remaining causes other than those represented by the indicators included in the model (Diamantopoulos, 2006, p. 11). Note that $x_1$ cannot be conceived as a »representative« indicator from a larger set of possible items because formative measurement does not follow the domain sampling model. This means that the items are not interchangeable (Rossiter, 2002, p. 315, original emphasis).11

Bearing the above in mind, a single-item measure under a formative perspective appears problematic for several reasons. First, because the latent variable is determined by its indicators rather than vice-versa, content specification of the construct is inextricably linked with indicator specification (Diamantopoulos and Winklhofer, 2001, p. 271). This implies that a census of indicators is needed (Bollen and Lennox, 1991). It is difficult to convincingly argue that such a census has been achieved by just using a single item. Second, as MacKenzie, Podsakoff and Jarvis (2005, p. 727) point out, »it is not conceptually possible for a formative indicator to perfectly represent a composite latent construct because the construct is defined as function of multiple distinct components

---

7 As indicated, for example, by the item’s squared multiple correlation (SMC) in a LISREL analysis (see Bollen, 1989).
8 In fact if the chosen item has much higher reliability and/or loading than the rest of the items, one could argue that it would not be representative of the other items!
9 It is assumed that $E(\varepsilon_1) = 0$ and $\text{COV}(\eta_1, \varepsilon_1) = \text{COV}(\eta_2, \varepsilon_2) = 0$.
10 For measurement models with multidimensional items, see Bollen (1989, pp. 198–199).
11 The authors would like to thank an anonymous reviewer for suggesting expanding the discussion of single-item reflective measures to models where item unidimensionality is not assumed.
12 Note that, like all formative models, the model shown in equation (4) is, on its own, statistically underidentified. Note also that addition of more items does not overcome the identification problem and »such models are unidentified unless embedded within a larger model which includes additional constructs measured with reflective indicators… or unless additional reflective items are included as consequences of the formative constructs (Diamantopoulos, 2006). An excellent discussion of identification issues in the context of formative models can be found in Tenen (2006).
or parts, so one indicator cannot validly represent the entire conceptual domain. Third, with a single indicator, it is more than likely that the variance of the residual (\(\text{VAR}(C)\)) in equation (4)) will be substantial (Diamantopoulos, 2006). In this context, as the residual increases, the meaning of the construct becomes progressively ambiguous... Certainly, the meaning of a construct is elusive when most of its variance is attributable to unknown factors (Williams, Edwards and Vandenberg, 2003, p. 908).\(^{12}\)

There is, however, one way one can conceive a single-item measure within a formative measurement perspective, namely, to specify a global item that summarizes the essence of the construct that the formative index purports to measure (Diamantopoulos and Winklofer, 2001, p. 272). Note that this item is not part of the set of formative indicators used to specify the focal construct but is typically used as an initial check of the validity of the individual indicators (e.g., see Diamantopoulos and Siguaw, 2006). A good example of this approach can be found in the job satisfaction literature where a global or overall single-item measures of satisfaction are compared with multi-item scales capturing different facets of job satisfaction (e.g., see Wanous, Reichers and Hudy, 1997; Oshagbemi, 1999; and Dolbier et al, 2005 and references therein). A theoretical rationale for using such global job satisfaction measures is that a worker would be generally satisfied if s/he is satisfied in all important facets of job satisfaction (e.g., satisfaction with pay, with supervisor, with co-workers, with promotion and with work itself...). In other words, the facet satisfaction leads to overall satisfaction (Law and Wong, 1999, p. 149). Such a global item approach to specifying single-item measures has been used in several other research fields, including quality of life assessment (e.g., de Boer et al, 2004), sport management (e.g., Kwon and Trail, 2005), citizen satisfaction (e.g., Van Ryzin, 2004), self esteem (Robins, Hendin and Trzesniewski, 2001), and teaching effectiveness assessment (e.g., Wanous and Hudy, 2001).

It is important to note, however, that none of the aforementioned studies (either in the job satisfaction or other research fields noted above) explicitly used a formative perspective as a theoretical foundation when specifying a global or overall single-item measures. The latter have been invariably developed on an ad hoc basis and, in fact, the multi-item measures with which they have been compared are typically reflective scales! This raises issues of measurement model misspecification for several of the constructs examined (job satisfaction being a prime example).\(^{13}\) Although such misspecification issues are clearly beyond the scope of the present paper, it is important to bear in mind that they do exist and suggest caution in interpreting the relevant literature.\(^{14}\)

In conclusion, there are several thorny issues – primarily of a conceptual nature – associated with the specification and selection of single-item measures under both a reflective and formative measurement perspective. Unfortunately, these issues are hardly ever addressed by proponents of single-item measures; Rossiter (2002, p. 314, added emphasis), for example, states that the goal is to develop one good item for each first-order component (meaning each facet of a formatively-measured construct). However, he offers no insights as to exactly what makes a good item (which is clearly what the real issue is all about). In this context, the typical approach in relevant methodological research has been to infer the quality of a single-item measure (usually specified on a more or less ad hoc basis) by comparing its relative performance with an established (and psychometrically sound) multi-item scale of the focal construct in terms of reliability and/or predictive validity. The underlying rationale is well-described by Gardner et al (1998, p. 899) who state that it is possible that one good item can be better than many bad items when evaluated on criteria of reliability and validity. It is to a discussion of such issues that we now turn.

3. Reliability Assessment

Reliability, in the context of measurement, is defined as the degree to which measures are free from error and therefore yield consistent results.
(Peter, 1979, p. 6). Reliability constitutes a crucial psychometric premise of any measure (Spector, 1992; Netemeyer, Bearden and Sharma, 2003). It is commonly agreed that longer scales are more reliable than short scales (Nunnally and Bernstein, 1994; Spector, 1992; Lowenthal, 2001; Rust and Golombek, 1989). Single-item measures in particular, are deemed to be «notoriously unreliable» (Spector, 1992, p. 4) both because measurement errors of the individual items are not smoothed out by the summation of the item-scores to a total score and because inconsistent responses may be obtained in successive administrations of the research instrument (Churchill, 1979; Epstein 1979). On the other hand, there is evidence to suggest that «additional items can significantly inflate error term correlation and thereby reduce the informal value of added items» (Drolet and Morrison, 2001, p. 200) meaning that the incremental information provided by each additional scale item is extremely small. Specifically, if items are semantically similar, respondents tend to make inferences from the content of one item to the remaining items of the scale, assuming that they are basically the same without carefully reading them. In general, respondents who are exposed to more items tend to distinguish less between them, with earlier items having a stronger influence on later items; hence more items may lead to mindless response behaviour (Drolet and Morrison, 2001). A second related problem is that multiple items are prone to consistency motif bias, «in which subjects tend to try to maintain consistency in their responses of similar questions» (Podsakoff et al., 2003, p. 881).

However, whether a single-item or multi-item measure is more reliable is, ultimately, an empirical question. Concerning test-retest reliability, there is evidence suggesting that single-items can be very reliable (e.g., see de Boer et al, 2004; Shamir and Kark, 2004). As far as internal consistency reliability is concerned, again, there is substantial evidence indicating acceptable reliability values for single-item scales (e.g., see Ginns and Barrie, 2004; Wanous, Reichers and Hudy, 1997; Wanous and Hudy, 2001; Kwon and Trail, 2005; Jordan and Turner, 2005; Dolbier et al, 2005). Note, in this context, that it is frequently said that one cannot estimate the internal consistency reliability of single-item measures and this alone is sometimes believed to be a sufficient reason to limit or avoid their use» (Wanous, Reichers and Hudy, 1997, p. 247). This misconception is particularly prevalent among management researchers as illustrated by the following passage from a review paper in the highly prestigious Strategic Management Journal:

«Single measures at the nadir of methodological sophistication, provide the researcher with the least assurance that a measure is a valid and reliable proxy of a construct and no estimates of reliability, and thus error, are possible» (Boyd, Gove and Hitt, 2005, p. 244).

Such misconceptions can probably be traced to the fact that, in structural equation models with reflective indicators, no parameter estimation can take place when only a single indicator is used to operationalize the construct of interest; this, however, is simply because the relevant measurement model is underidentified (see earlier discussion of equation (1)) and it makes no difference as to which particular indicator is used in the model. When a single-item measure is contrasted with a multi-item measure, however, there are several ways of establishing the reliability of the former based on such techniques as factor analysis and the correction for attenuation formula (see Wanous and Reichers, 1996 and Wanous and Hudy, 2001 for formulae and illustrative applications). Moreover, assuming that one wants to select a single item from a scale of \( k \) items, one can apply the Spearman-Brown prophecy formula in reverse (see Nunnally and Bernstein, 1994, p. 263–264), whereby the reliability of a scale «1/\( k \) th» as long as the original scale would be estimated.

In summary, an outright rejection of single-item measures on reliability considerations does not seem to be justified. One can estimate the reliability of single-item measures and the resultant estimates are, more often than not, within acceptable levels. However, «reliability is a necessary but not sufficient condition for validity» (Churchill, 1979, p. 65; see also Peter, 1979). A discussion of validity

---

16 Note, however that «an increase in the number of the scale items leads to participant fatigue, boredom, and inattention, which, in turn, can lead to inappropriate [mindless] response behaviours» (Drolet and Morrison, 2001, p. 198). Hence, there is a trade off between the parsimony of a scale and its reliability (DeVellis, 2003).

15 Of course, there is also evidence to the contrary, that is, indicating insufficient reliability of single-item scales (e.g., see Loo, 2002; Epstein, 1979).
Given that case, then obviously »one or two than one conceptual domain. If this is indeed the certained that long scales possessed sound reliabil-
(1995), who, in a review of scale devel-
ally decrease the content validity of the measure« (Peter, p. 331). This problem is confirmed from different angles (Baumgartner and Homburg, 1996; Wirtz and Lee, 2003). On the other hand, these arguments are traded off by the (common) »practice of adding attempted synonyms that actu-
ity revealed mean observed and corrected corre-
loadings λs and λ3 or of discriminant validity (as captured by the magnitude of the interconstruct correlation, Φ) can be obtained. Thus when both constructs are measured with just one indicator each, it is not possible to either distinguish between the constructs or test whether each indicator in-
deed loads stronger on its posited construct rather than on a different construct. For instance, the constructs of »brand attitude« and »brand satisfac-
tion« would not be formally distinguishable when single-item measures are used.

In light of the above, the common approach for assessing convergent validity has been the com-

tion« would not be formally distinguishable when single-item measures are used.

4. Validity Assessment

Construct validity refers to »the degree to which a measure assesses the construct it is purported to assess ... In this sense a measure is construct valid (1) to the degree that it assesses the magnitude and direction of representative sample of the character-
istics of the construct and (2) to the degree that the measure is not contaminated with elements from the domain of other constructs and error« (Peter, 1981, p. 134).

It is asserted that single-item measures lack va-

lidity, because they tend to insufficiently capture the conceptual domain of most constructs (Nun-
nally and Bernstein, 1994); thus compared to multi-
item measures, they do not tap into a construct from different angles (Baumgartner and Homburg, 1996; Wirtz and Lee, 2003). On the other hand, these arguments are traded off by the (common) »practice of adding attempted synonyms that actu-
ify than multiple, poorly presented items that increase the error term correlations and/ or stimulate inappropriate response styles« (Drolet and Morrison, 2001, p. 199, original emphasis). A related point, from a face validity perspective, is that single-item measures may be advantageous because »respondents may resent being asked ques-
tions that appear to be repetitions... From a man-
agement perspective, a single-item is usually easier to understand than a scale score« (Wanous, Reichers and Hudy, 1997, p. 250).

A particular concern with single-item scales concerns the assessment of convergent and discrimi-

nantly valid items that each item tends to relate to attributes other than the one to be measured« (Nunnally and Bernstein, 1994, p. 66). Given two constructs η1 and η2 each measured by a single in-
dicator, say, x1 and x2 respectively and assuming standardization of the latent variables, five param-
eters need to be estimated, namely λ11, λ22, θ1 = VAR(ε1), θ2 = VAR(ε2) and Φ12 = COV(η1, η2).17 As there are only three pieces of information available (i.e., VAR(x1), VAR(x2) and COV(x1, x2)), the model is underidentified and, therefore, no estimates of convergent validity (as captured by the indicator loadings λ1 and λ3) or of discriminant validity (as captured by the magnitude of the interconstruct correlation, Φ) can be obtained. Thus when both constructs are measured with just one indicator each, it is not possible to either distinguish between the constructs or test whether each indicator in-
deed loads stronger on its posited construct rather than on a different construct. For instance, the constructs of »brand attitude« and »brand satisfac-
tion« would not be formally distinguishable when single-item measures are used.

In light of the above, the common approach for assessing convergent validity has been the com-
putation of the correlation between the single-item measure and its multi-item counterpart (i.e., the »full« measure of the construct). The empirical evi-
dence based on this approach has been, overall, encouraging for single-item measures. For exam-
ple, in a job satisfaction context, Wanous, Reichers and Hudy’s (1997) meta-analytic study reported a mean correlation between single items and full scales of .63 and a corrected correlation of .67.18

Another meta-analytic study of teaching effective-
ness revealed mean observed and corrected corre-
lations of .79 and .84 respectively (Wanous and Hudy, 2001). Similarly encouraging results have been reported by Robins, Heudin and Trzesniewski (2001), Nagy (2002), and Dolbier et al. (2005), among others. Although the discriminant validity of single-item measures has been investigated less often, those (few) studies who have done so also reported findings in support of single-item mea-
ures (e.g., see Gardner et al, 1998, and de Boer et al., 2004).

Finally, several studies have compared the pre-
dictive validity of single-item measures to that of multi-item scales. With few exceptions, the evi-
dence from studies in fields as diverse as health care (DeSalvo et al, 2006), sports management (Kwon and Trail, 2005), organizational psychology

17 Given that η1 and η2 are standardized, Φ is a correlation coefficient.
18 Corrected for unreliability.
(Nagy, 2002) and marketing (Bergkvist and Rossiter, 2007) shows that single-item scales can have good predictive validity (comparable to those of their multi-item equivalents).

Taken collectively, the findings relating to the validity of single-item measures fail to support the classic psychometric argument (e.g., Churchill, 1979; Nunnally and Bernstein, 1994) that multiple-item measures are more valid than single-item measures for all types of constructs* (Bergkvist and Rossiter, 2007, p.182). Thus single-item measures can be both reliable and valid, implying that their potential application should not be ruled out on a priori grounds.

5. Criteria for the Use of Single-Item Measures

In the previous two sections it was shown that, contrary to common beliefs, single-item measures can have acceptable psychometric properties. In this context, it has been argued that «if single-item measures can be shown to adequately and accurately represent the variables being measured, then many researchers ... could benefit. There are several advantages to using single-item scales: simplicity, brevity or ease of use, and global measurement» (Kwon and Trail, 2005, p. 72). We therefore now establish potential criteria that should assist researchers decide whether the use of single-item measures is acceptable or not. In doing so, we identify conditions or situations, in which the use of single items would be worth serious consideration. The focal criteria can be summarized into four broad groups corresponding to (1) the nature of the construct, (2) the nature of existing instruments, (3) the research objectives, and (4) sampling considerations.

5.1. Nature of the Construct

The selection of single- versus multiple-items depends to a great extent on the construct of interest. Particularly relevant in this respect is whether the focal construct is concrete or abstract (Rossiter, 2002). Concrete constructs refer to objects and their characteristics which are perceived similarly by all raters, whereby there is «virtually unanimous agreement by raters as to what it is, and they clearly understand that there is only one or holistically one, characteristic being referred to when the attribute is posed, as in a questionnaire or interview, in the context of the to-be-rated object» (Rossiter, 2002, p. 313). Examples are favorability, price perception, and buying intention.

Abstract constructs, on the other hand, mean different things to different raters and thus, are perceived as heterogeneous by raters; examples are creativity, power, or corporate culture. It is generally held that, with abstract constructs, the use of multiple-item measures is required, because «most constructs, by definition, are too complex to be measured effectively with a single item...» (Peter, 1979, p. 16). On the other hand, when a construct is judged to be concrete, the use of single item measures is considered reasonable (Sackett and Larson, 1990; Rossiter, 2002), not least because measurement error is more prevalent for abstract versus concrete concepts (Cote and Buckley, 1987). We basically share this view and advocate the use of single-item measures for investigating concrete constructs.

Related to the concreteness of the construct is its complexity/dimensionality. Generally, for multidimensional constructs, the use of single-item measures is inappropriate (Drolet and Morrison, 2001b, Nunnally and Bernstein, 1994). According to Loo (2002, p. 73), for example, «single-item measures may be considered only if the single item reflects a homogeneous construct, as indicated by a high internal consistency reliability coefficient (alpha > 0.85) or a unidimensional construct as reflected by an item factor analysis.» Hence, the use of single-item measures is appropriate if the construct of interest is unidimensional rather than multidimensional (Sackett and Larson, 1990; Nunnally, 1978). However, in cases in which a construct is multidimensional and its respective dimensions are known and have been empirically established, each dimension may be measured with a single item, provided that the latter is reliable.

While it is usually the case that with increasing complexity of the construct (i.e., increasing number of dimensions), additional questions (and hence items) need to be asked to cover all dimensions of the latter, «there comes a point where the construct becomes so complex that a single question may be

19 See also the discussion in Section 2.2 regarding item dimensionality.
Using single-item measures for construct measurement

the best.« (Sloan et al., 2002, p. 481). With *highly* complex constructs, it may be that not all dimensions of the latter are covered by traditional multi-item measures, thus, resulting in incomplete evaluations of the construct (Nagy, 2002). Put differently, in order to cover each and every potential aspect of a highly complex construct, an enormous number of items would be needed that would render a practical application impossible. As a result, Scarpello and Campbell (1983), for example, contend that the best way to measure overall job satisfaction is to use a single question along the lines »Overall, how satisfied are you with your job?«

When asking »global« single-item questions such as the above, it is assumed that respondents *automatically* consider different aspects of the construct (provided that respondents are aware of the full scope of the concept). Single-item measures allow a respondent to *consider all aspects and individual preferences of the certain aspects of the construct being measured* (Nagy, 2002, p. 79), and thus provide a more »tailor-made« picture of that respondent’s construct view. Consequently, with »global« single-item measures, respondents tend to ignore aspects that are not relevant to their situations, and differentially weight the relevant aspects to provide a single rating (De Boer et al., 2004); in contrast, multiple-item scales employ external subjective or statistical weighting to combine items to come to an overall rating. In this context, the question arises whether it is better if researchers decide which facets of the concept are to be measured, and weigh them based on statistical considerations or if consumers should be empowered to decide which facets are important or not important. For example, if consumers are asked for their quality of life status, they may consider more facets or alternative facets which may be of particular importance to them, compared to a fixed number of facets included in standardized multiple-item measures.

### 5.2. Nature of Existing Instruments

As already mentioned in previous sections, in an effort to maximize the internal consistency of scales, numerous constructs in business and management research possess semantically identical and therefore redundant items (Albers and Hildebrandt, 2006), whereby »essentially the same item is rephrased in several different ways« (Boyle, 1991, p. 281). With semantically identical items one aspect of the domain is likely to be oversampled (Smith and McCarthy, 1995). This tends to lead to the effect that the same error variance associated with the particular item is also likely to be associated with the other items (Drolet and Morrison, 2001, Smith and McCarthy, 1995); as a result, measurement error does *not* cancel out in the aggregation process, which negatively affects reliability and precision.

Redundant items also contribute to the known negative effects of lengthy scales, namely boredom, fatigue, and various response biases (Duchachek, Coughlan, and Iacobucci, 2005; Drolet and Morrison, 2001). Hence, we advise researchers to carefully scrutinize existing scales and determine the extent to which their items are semantically identical/similar; this judgment should be conducted by at least two experts (coders) independently to ensure a high degree of objectivity. Very high estimates of item homogeneity (as reflected in inter-item correlations) are indicative of potential

---

20 Thus single-item measures »allow the subject to take personally salient features of the situation into account when providing a response« (Youngblut and Casper, 1993, p. 459).

21 As pointed out by Sloan et al., 2002, p. 484, »the level of abstract thinking required of the respondent may be greater when using single-item measures than when using multiple-item indices: For example, it may be more cognitively challenging to provide an overall rating of one's level of social functioning than to respond to a series of relatively concrete questions about spending time with family and friends«.
item redundancy (Boyle, 1991). Note, however, that high inter-item correlations should not be automatically equated with semantically redundant items, because items measuring different aspects of a construct may be highly related. However, in case where a high degree of redundancy among items is established, we subscribe to Drolet and Morrison’s (2001, p.197) view that »information content can be negatively affected by scales with multiple items« and, therefore, strongly recommend the use of a single-item measure.

5.3. Research Objectives

The issue whether to use a single-item measure also depends on the respective objective(s) of the study, and, in particular, the role of the construct in the research design (Sloan et al., 2002). If a construct is at the heart of a study, with the researcher’s intention to generate specific insights into the nature of the construct, a more detailed measurement approach applying full-length multiple-item scales should be chosen. The latter are superior over single-item measures for identifying detailed aspects or highlighting certain characteristics of constructs (Lee et al., 2000). If, on the other hand, investigators are only interested in obtaining a general view of the construct and the research objective is to get an overall feeling, judgment, or impression on the latter, a single-item measure is often adequate for the purpose (Poon, Leung and Lee, 2002). Indeed, »the single-item global rating method may be useful if the goal of a study is to gain an understanding for the general nature of a construct« (Lee et al., 2000, p. 242). Finally, if the construct is only of secondary importance in the study setting (e.g., is used as moderator, validation, or control variable), the employment of single-item measures can be justifiable.22

In contrast to multi-item measures, single-item measures are criticized for lacking precision because they have a tendency to categorize people into a relatively small number of groups (e.g. Spector, 1992; DeVellis, 2003; Nunnally and Bernstein, 1994; Netemeyer, Bearden and Sharma, 2003; Churchill, 1979). Hence, multiple item measures possess more responsiveness, that is, the ability of a measure to detect small but important differences (de Boer et al., 2004).23 This is crucial, because many measurement problems require a very fine differentiation or categorization (Nunnally and Bernstein, 1994) and some measures are purposely constructed with the aim to clearly differentiate among respondents or to differentiate extreme cases from less extreme ones (Smith and McCarthy, 1995). Having said all that, the inclusion of more points in a single-item measure may be a (partial) remedy for the precision issue (Bergkvist and Ros-siter, 2007). For example, according to Wirtz and Lee (2003), more finely grained single-item measures are able to discriminate between different degrees of the measure similarly to multiple-item measures.

In longitudinal studies, the use of single-item measures is often recommended over multi-item measures because regularly completing a lengthy scale (or questionnaire) puts a big burden on the respondent (Drolet and Morrison, 2001; Nagy, 2002; Wanous, Reichers and Hudy, 1997). However, although well-meant, such recommendations may be misguided because with a single-item measure, it is not possible to differentiate between (a) »true« changes in the underlying construct over time, (b) reliability of measurement, and (c) situational factors that may impact responses when measurement takes place at different points in time (as is the case in longitudinal studies). Indeed, according to latent state-trait theory,24 »measurement does not take place in a situational vacuum; a person can be measured only in a situation which might have a systematic effect on each variable measured at occasion k ... Several instruments measuring the same state on each occasion are necessary in order to obtain information about the reliability of the measurements. Repeated observations on several occasions are necessary in order to get information on the degree of consistency over interindividual differences over time« (Steyer and Schmitt, 1990, p. 433 and p. 437, emphasis in original).

In light of the above, we advise against the use of single-item measures in longitudinal research

22 For example, in a strategic management setting, Boyd, Gove and Hitt (2005) show that control variables are nearly exclusively captured using single items (79.8% of control variables); besides, 38.1% of dependent variables and 47.6% of independent variables used single-item indicators.

23 In this context, the relatively large number of total scores makes it possible to »make relatively fine distinctions among people (Churchill, 1978, p. 66).

24 For a discussion of latent state-trait models and applications, see Steyer and Schmitt (1990), Steyer, Ferring and Schmitt (1992) and Steyer, Schmitt and Eid (1999).
and subscribe to Steyer’s Schmitt and Eid’s (1999, p. 403) recommendation that »we need repeated measurements on at least two occasions with at least two instruments measuring the same states«.

5.4. Sampling considerations

If a measure is to be administered to a wide range of different populations, the use of single-item measures has certain advantages. Indeed, an »advantage of the single-item or short scales is that they can be given to numerous people« (Gorsuch and McPherson, 1989, p. 352). Indeed, one major strength of single-item measures is their flexibility; for example, »whether one is investigating the job satisfaction of pilots, fishermen, solicitors or managers, the single-item measure of job satisfaction may be administered« (Oshagbemi, 1999, p. 393).

In the field of organizational research, managers frequently feel that they are »over-surveyed«, which contributes to the issue of low response rates (Rogelberg and Luong, 1998). Specifically, the difficulty of attaining large sample sizes in surveys, due to lack of willingness to sacrifice time to complete questionnaires, leads to the necessity of reducing the length of construct measures. As a rule of thumb, there should be at least ten times as many respondents as items or, in cases where a large number of items are used, at least five respondents per item (Nunnally, 1967; Peter, 1979).

Taking these considerations into account, we suggest that if only a small sample size is available (e.g., due to budget constraints, difficulties in recruiting respondents, requirement of dyadic data, and other similar constraints), considering the use of single-item measures may be a pragmatic solution.

Table 1 summarizes the criteria along which researchers may judge whether measurement with a single item is acceptable/justifiable. The more criteria that are fulfilled, the more acceptable the use of single-item measures. For instance, if a construct is concrete, used as a control variable, and the single-item measure sufficiently satisfies reliability and validity requirements, it would indeed be appropriate to use it in substantive research.

6. Conclusions

This article contributes to the management literature by providing insights into the conceptual nature of single-item measures and by dispelling several myths regarding their psychometric properties. Specifically, our analysis shows that the application of single-item measures is appropriate under certain conditions and that their general banishment is not justified. By considering the criteria offered in this paper, researchers should be able to reach an informed decision as to whether the use of a single-item measure for the focal construct is acceptable given the research purpose and setting at hand. Assuming that it is, it is important to select/

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Multi-item scale</th>
<th>Single-item scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct concreteness</td>
<td>Abstract</td>
<td>Concrete</td>
</tr>
<tr>
<td>Construct dimensionality/complexity</td>
<td>Multidimensional, moderately complex</td>
<td>Unidimensional or extremely complex</td>
</tr>
<tr>
<td>Semantic redundancy</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Primary role of construct</td>
<td>Dependent or independent variable</td>
<td>Moderator or control variable</td>
</tr>
<tr>
<td>Desired precision</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Monitoring changes</td>
<td>Appropriate</td>
<td>Problematic</td>
</tr>
<tr>
<td>Sampled population</td>
<td>Homogenous</td>
<td>Diverse</td>
</tr>
<tr>
<td>Sample size</td>
<td>Large</td>
<td>Limited</td>
</tr>
</tbody>
</table>

Table 1: Criteria for Assessing the Potential Use of Single-Item Measures
develop a single-item measure with acceptable psychometric properties. In this context, a difficulty facing the use of single-item scales is the construction of the single-item measures to adequately represent the desired construct. These measures must be constructed carefully to achieve the desired reliability and to comply with common variance in the dependent variable (Kwon and Trail, 2005, p. 84). Unfortunately, measurement theory has only paid scant attention to systematic methods for developing and validating single-item measures. This is a gap in need of future research and for which strategies of scale reduction (e.g., see Stanton et al., 2002) or (parts of) Rossiter’s (2002) C-OAR-SE scale development procedure may represent useful starting points; the use of item response theory (e.g., see Hambleton and Swann-nathan, 1985) is another option worth pursuing as recently suggested by Salzberger (2007).

Relating to the above point, there is a need to document empirical findings on the psychometric qualities of single-item measures in scale handbooks (e.g., see Bearden and Netemeyer, 1999) which should facilitate an increased diffusion and higher acceptance of single-item measures, and would represent a strong argument in favor of the latter in the review process of publications. Furthermore, we suggest that authors developing multi-item scales should, if appropriate, also test and report single-item versions of the latter. This would enable researchers, depending on the detail of the analyses demanded, to choose either a single-item or a full-length version, and give investigators additional confidence in the psychometric soundness of single-item measures.

Finally, it is worth recalling that current measurement practice in management is largely borrowed from ability-test theory in psychology, where the items differ in difficulty and there is within-person variation in ability to answer them... For ability tests a single item cannot provide a precise (reliable) estimate of the individual’s ability (Rossiter, 2002, p. 321, original emphasis). Thus the question arises if we really need highly sophisticated multi-item measures for all business-related constructs and treat the latter as if they were psychological (personality and behavioral) and therefore potentially much more complex constructs.25 Basically, there is no doubt that complex (soft) constructs in psychology like depression or anxiety require the need of sound multi-item measurement, but is this also the case with constructs like purchase likelihood or overall satisfaction with a brand? For the latter, it is entirely conceivable that theoretical tests and empirical findings would be unchanged if good single-item measures were substituted for these constructs in place of commonly used multiple-item measures (Bergkvist and Rossiter, 2007, p. 183).

In conclusion, it may well be time to challenge conventional measurement wisdom in management research concerning the (non-) use of single-item measures. While the latter do not, in any way, present a sound measurement option under all circumstances, there are conditions under which they do and, therefore, the use of single-item measures should not be considered fatal flaws in the review process (Wanous, Reichers and Hudy, 1997, p. 270).

References


25 Generally, psychological assessment about a person or mood is much more complex and also yields more effects than for instance attitudinal measures, all which are closely related. Hence, the complete transfer of psychological measurement procedures should not be fully transferred to measurement in business strategy, since these are two distinct domains.

Christoph Fuchs/Adamantios Diamantopoulos

DBW 69 (2009) 2


Law, Kenneth/Wong, Chi-Sum (1999): Multidimensional con-


