J. of the Acad. Mark. Sci. (2012) 40:434–449
DOI 10.1007/s11747-011-0300-3

METHODOLOGICAL PAPER

Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective

Adamantios Diamantopoulos · Marko Sarstedt · Christoph Fuchs · Petra Wilczynski · Sebastian Kaiser

Received: 1 June 2011 / Accepted: 27 December 2011 / Published online: 14 February 2012

Abstract Establishing predictive validity of measures is a major concern in marketing research. This paper investigates the conditions favoring the use of single items versus multi-item scales in terms of predictive validity. A series of complementary studies reveals that the predictive validity of single items varies considerably across different (concrete) constructs and stimuli objects. In an attempt to explain the observed instability, a comprehensive simulation study is conducted aimed at identifying the influence of different factors on the predictive validity of single versus multi-item measures. These include the average inter-item correlations in the predictor and criterion constructs, the number of items measuring these constructs, as well as the correlation patterns of multiple and single items between the predictor and criterion constructs. The simulation results show that, under most conditions typically encountered in practical applications, multi-item scales clearly outperform single items in terms of predictive validity. Only under very specific conditions do single items perform equally well as multi-item scales. Therefore, the use of single-item measures in empirical research should be approached with caution, and the use of such measures should be limited to special circumstances.

Keywords Single items · Multi-item scales · Predictive validity · Measurement theory

The authors thank Edward E. Rigdon (Georgia State University), Udo Wagner (University of Vienna) and the anonymous reviewers for their helpful comments on previous versions of this paper.

A. Diamantopoulos
Department of Business Studies, University of Vienna,
Bruenner Strasse 72,
1210 Vienna, Austria
e-mail: adamantios.diamantopoulos@univie.ac.at

M. Sarstedt (✉)
Institute for Market-based Management,
Ludwig-Maximilians-University Munich,
Kaulbachstrasse 45,
80539 Munich, Germany
e-mail: sarstedt@bwl.lmu.de

M. Sarstedt
Faculty of Business and Law, University of Newcastle,
Newcastle, Australia

C. Fuchs
Rotterdam School of Management,
Erasmus University,
Burgemeester Oudlaan 50,
3062 PA Rotterdam, The Netherlands
e-mail: cfuchs@rsm.nl

P. Wilczynski
Institute for Market-based Management,
Ludwig-Maximilians-University Munich,
Kaulbachstrasse 45,
80539 Munich, Germany
e-mail: wilczynski@bwl.lmu.de

S. Kaiser
RSU Rating,
Karlstrasse 35,
80333 Munich, Germany
e-mail: sebastian.kaiser@rsu-rating.de

Introduction

More 30 years ago, in a widely cited *Journal of Marketing* article, Jacoby (1978, p. 93) alerted researchers to the "Folly of Single Indicants," arguing that "given the complexity of our subject matter, what makes us think that we can use responses to single items […] as measures of these concepts, then relate these scores to a host of other variables, arrive at conclusions based on such an investigation, and get away calling what we have done 'quality research'?" Marketing academia was quick to respond to Jacoby's (1978) criticism as evidenced in a series of highly influential papers seeking to provide guidance to researchers in their measure development efforts (e.g., Churchill 1979; Churchill and Peter 1984; Peter 1979). The adoption of structural equation modeling techniques further encouraged the systematic psychometric assessment of multi-item (MI) scales in terms of dimensionality, reliability, and validity (e.g., Anderson and Gerbing 1982; Steenkamp and van Trijp 1991). Nowadays, the use of MI scales is standard practice in academic marketing research as reflected both in relevant scale development monographs (e.g., Netemeyer et al. 2003; Viswanathan 2005) and in numerous handbooks containing compilations of marketing measures (e.g., Bearden et al. 2011; Bruner et al. 2005).

Recently, however, Bergkvist and Rossiter (2007, p. 183) challenged this conventional wisdom on both theoretical and empirical grounds and concluded that "theoretical tests and empirical findings would be unchanged if good single-item measures were substituted in place of commonly used multiple-item measures." Their theoretical challenge was based on the C-OAR-SE procedure (Rossiter 2002, 2011) according to which, if the object of the construct (e.g., a brand or an ad) can be conceptualized as concrete and singular and if the attribute of the construct (e.g., an attitude or a perception) can be designated as concrete, there is no need to use an MI scale to operationalize the construct. Furthermore, Bergkvist and Rossiter (2007, 2009) reported empirical findings indicating that single-item (SI) measures demonstrated equally high predictive validity as MI scales. The authors' final conclusion was that "carefully crafted single-item measures—of doubly concrete constructs—are at least as valid as multi-item measures of the same constructs, and that the use of multiple items to measure them is unnecessary" (Bergkvist and Rossiter 2009, p. 618).

In light of Bergkvist and Rossiter's (2007, 2009) findings, researchers may be tempted to adopt SI measures not least because the latter have numerous practical advantages such as parsimony and ease of administration (e.g., Drolet and Morrison 2001; Fuchs and Diamantopoulos 2009; Wanous et al. 1997). Given recent concerns regarding "oversurveying," decreasing response rates, and high costs of surveying additional items (Rogelberg and Stanton 2007), the adoption of SI measures is clearly tempting. However, caution needs to be exercised before established MI scales are abandoned in favor of single items, for at least three reasons.

First, research in other fields shows that SI measures do not *always* perform as well as MI scales of the same construct. For example, in a study by Kwon and Trail (2005), sometimes the MI scale outperformed the SI measure, sometimes there was no difference between them, and sometimes the SI measure was a better predictor than the MI scale. Overall, the results varied both across constructs and depending upon the specific criterion variable under consideration (see also Gardner et al. 1989; Loo 2002).

Second, the response pattern of an item measuring a specific construct frequently carries over to the subsequent item measuring (the same or) another construct due to respondents' state dependence (De Jong et al. 2010). If the subsequent item is the *only* item measuring another construct (i.e., an SI measure), such carry-over effects might considerably affect the measure's (predictive) validity. The use of multiple items, in contrast, may compensate such effects.

Third, prior studies (Bergkvist and Rossiter 2007, 2009) have used Fisher's z-transformation test to compare correlation coefficients and R^2-values when contrasting the predictive validity of SI versus MI measures. However, this test is only appropriate when correlations from two *independent* (as opposed to paired) samples are to be compared (e.g., Steiger 1980); for related correlation coefficients, Ferguson's (1971) or Meng et al.'s (1992) procedures should be employed.

Given the practical advantages of SI measures, evidence legitimating their use is clearly welcome. At the same time, evidence to the contrary cannot be ignored either. Against this background, the present study seeks to investigate under which conditions SI measures are likely to have comparable predictive ability as MI scales. We first replicate Bergkvist and Rossiter's (2007, 2009) analyses by comparing the predictive validity of SI and MI measures of attitude toward the ad (A_{Ad}) and attitude toward the brand (A_{Brand}). We then undertake a similar analysis using different (concrete) constructs to ascertain the robustness of our findings in different settings. We find evidence indicating that SI measures *can* have predictive ability similar to MI scales; however, we also observe that the latter significantly outperform the former in most of our empirical settings. Thus, whereas a particular SI may yield good results in one setting (e.g., in one product category), the same item's predictive validity may be disappointing in another.

To shed light on the observed instability, we subsequently conduct a simulation study to identify the influence of different design characteristics (e.g., the average inter-item correlation among the items of the predictor and criterion constructs, the number of items used to measure these constructs) on the predictive validity of SI versus MI measures. By systematically varying different combinations of these characteristics, we offer insights into the relative performance of SI and MI scales under different conditions. Based on our findings, we then provide marketing researchers with an empirically-based guideline for the use of SI and MI scales in practical applications.

Theoretical background

According to conventional measurement theory, the (reflective) items comprising an MI measure of a focal construct represent a random selection from the hypothetical domain of all possible indicators of the construct (Nunnally and Bernstein 1994). Using multiple items helps to average out errors and specificities that are inherent in single items, thus leading to increased reliability and construct validity (DeVellis 2003). In this context, "in valid measures, items should have a common core (which increases reliability) but should also contribute some unique variance which is not tapped by other items" (Churchill and Peter 1984, p. 367). In practice, however, scale developers often place undue emphasis on attaining high reliability, resulting in semantically redundant items that adversely affect the measure's validity (Drolet and Morrison 2001). It is against this background that proponents of SI measures argue that "when an attribute is judged to be concrete, there is no need to use more than a single item […] to measure it in the scale" (Rossiter 2002, p. 313).

Although the above recommendation is undoubtedly appealing from a pragmatic point of view, it is not without problems from a conceptual perspective. Formally, given a single measure x_1 and an underlying latent variable η (representing the focal construct), the relevant measurement model is described by the following equation, where λ_1 is the loading of x_1 on η and ε_1 is measurement error, with $COV(\eta, \varepsilon_1)=0$ and $E(\varepsilon_1)=0$.

$$x_1 = \lambda_1\eta + \varepsilon_1, \qquad (1)$$

There are two possible ways of interpreting x_1 in Eq. 1, namely that (1) x_1 is somehow *unique* (i.e., no other item could possibly measure η) or (2) that x_1 is *representative* (i.e., it is interchangeable with other measures of η). The first interpretation is highly problematic because "an observable measure never fully exhausts everything that is meant by a construct" (Peter 1981, p. 134). Indeed, if x_1 where to be seen as *the* measure of η, "a concept becomes its measure and has no meaning beyond that measure" (Bagozzi 1982, p. 15). The second interpretation (x_1 as a representative measure of η) is more consistent with the domain sampling model but raises the question of how the item should be chosen. As Diamantopoulos (2005, p. 4) observes, "if … a single 'good' item is to be chosen from a set of potential candidates (which implies that *other* items could, in principle, have been used instead), the question becomes *how* to chose the 'best' (or at least, a 'good') item."

One option is to choose *a priori* one item from a set of indicators based on face validity considerations (e.g., Bergkvist and Rossiter 2007). However, given that *all* items in an MI scale should conform to the domain sampling model (DeVellis 2003; Nunnally and Bernstein 1994), there is no reason why any one item should be *conceptually* superior to the others; assuming unidimensionality, scale items are essentially interchangeable from a content validity point of view (Bollen and Lennox 1991).

Another option is to ask a panel of experts to select the item that "best captures" or "most closely represents" the focal construct (e.g., Rossiter 2002). This approach has the advantage that it is based on empirical agreement among expert judges rather than solely on the preferences of the researchers. However, the conceptual issue as to *why* the chosen item is better than the rest of the items is still not addressed. Also, there is evidence showing that experts are not infallible (Chi et al. 1988).

A third option is to base item choice on statistical criteria such as an indicator's communality (e.g., Loo 2002) or the reliability of the indicator (e.g., Wanous et al. 1997). While this approach explicitly considers the psychometric properties of the various scale items, it is also subject to sampling variability; for example, the item displaying the highest communality in one sample may not do so in another sample. Thus, identifying a suitable SI *prior* to statistical analysis is hardly feasible.

A fourth option is to choose an item at random. Random choice would appear to be most conceptually consistent with the domain sampling model. However, according to congeneric measurement (Jöreskog 1971), items may differ from one another both in terms of how strongly they relate to the underlying construct and in terms of their susceptibility to measurement error (Darden et al. 1984); thus random choice may or may not pick the "best" item (i.e., the item with the strongest loading or the smallest error variance).

A final option is to look outside the MI scale and generate a tailor-made SI measure (e.g., Bergkvist and Rossiter 2009). However, given the plethora of MI scales available for most marketing constructs, it is unclear what additional benefit would be gained by generating extra (i.e., "stand alone") SI measures. Moreover, there are no established procedures for the construction of SI measures in marketing.

In what follows, we contrast the predictive ability of MI scales against that of *each* individual item comprising the scales. Evidence in favor of using an SI would be provided if (1) *at least one* item displays comparable predictive validity as the entire scale, (2) the item(s) concerned does so across different samples, and (3) the item(s) concerned does so across different stimuli (e.g., brands or ads). The stability implied by (2) and (3) is essential because if SI performance is very variable in different settings, it becomes extremely difficult to *ex ante* select an SI as a measure of the focal construct in a planned study. Clearly, from a practical perspective, unless one can select a "good" item *before* the study is executed, the benefits of SI measures (e.g., parsimony, flexibility, less monotony, ease of administration) will not be reaped.

Study 1: replication of Bergkvist and Rossiter (2007, 2009)

Study 1 uses the same design, focal constructs and measures as Bergkvist and Rossiter (2007, 2009). Specifically, we compare the predictive validity of SI versus MI measures of attitude toward the ad (A_{Ad}), brand attitude (A_{Brand}), and purchase intention (PI_{Brand}) measured on 7-point semantic differential scales. We drew our data from a survey of 520 university students (age: M=22 years, 68% female) who were randomly exposed to two of four real advertisements of products in four different product categories: insurance, jeans, pain relievers, and coffee (Bergkvist and Rossiter 2007, 2009). The ads were taken from foreign countries to ensure that respondents knew neither the brands nor the ads.

We first confirmed the unidimensionality of the three MI scales using factor analysis and computed their internal consistencies, which were satisfactory (minimum α values of .87, .88, and .88 for A_{Ad}, A_{Brand}, and PI_{Brand}, respectively). We then computed the correlation (r) between the MI measures of A_{Ad} (predictor) and A_{Brand} (criterion) as well as between A_{Brand} (predictor) and PI_{Brand} (criterion). Next, we computed correlations between each individual item capturing A_{Ad} and the full A_{Brand} scale and compared the resulting correlation coefficient with that obtained in the previous step using Meng et al.'s (1992) test. We did the same for the items capturing A_{Brand} and the full PI_{Brand} scale. In line with measurement theorists (Bergkvist and Rossiter 2007, 2009; Carmines and Zeller 1979), we assume that the higher correlations, the closer they are to the true correlations (in the population). We also undertook a bootstrap analysis (Cooil et al. 1987; Efron 1979, 1981) to evaluate the relative performance of SI and MI measures in slightly changed data constellations. Table 1 summarizes the results.

The results relating to A_{Ad} (Table 1A-D) show that in three out of the four product categories, *all* individual items have significantly lower predictive validity than the full scale. Only for pain relievers there is a single instance (like/dislike) for which comparable performance is obtained with an SI. These findings are further supported by the bootstrapping results which show that, in the vast majority of sample runs, the MI scale outperforms the individual items.

A similar picture emerges for the relationship between A_{Brand} and PI_{Brand} (Table 1E-H). For example, good/bad displays a comparable predictive validity as the MI scale for pain relievers and coffee, but not for insurance and jeans. Similarly, pleasant/unpleasant performs equally well as the MI scale for pain relievers but not for any other product category; the other single items are consistently outperformed by their MI counterparts.

Our replication of Bergkvist and Rossiter (2007, 2009) thus reveals considerable variability in the performance of single items. Whereas, depending on the product category, some items have similar predictive validity as the MI scale, others consistently lag behind, suggesting that the relative performance of SI measures is context and construct-specific.[1] We further examine this issue using different constructs, different stimuli (brands), and non-students as respondents in Studies 2 and 3 below.

Study 2

Our second empirical study is based on a consumer sample and uses the hedonic (HED) and utilitarian (UT) dimensions of consumer attitudes towards products (Batra and Ahtola 1991) as focal constructs. Conceptually, the hedonic dimension measures the experiential enjoyment of a product, while the utilitarian dimension captures its practical functionality (Batra and Ahtola 1991; Okada 2005; Voss et al. 2003). We selected these constructs because, under Rossiter's (2002) terminology, each dimension can be considered as a doubly-concrete construct in that the object and the attribute of the construct "are easily and uniformly imagined" (Bergkvist and Rossiter 2007, p. 176); consumers are likely to easily understand the meaning of the items measuring the two constructs (e.g., enjoyable, useful) as a set of expert raters also confirmed. Previous applications of the HED and UT scales have produced alphas above .80 (Voss et al. 2003), and have even substituted the dimensions with single items (Okada 2005).

We used Voss et al.'s (2003) scales to capture the two dimensions (see Table 2) and a four-item measure of brand liking (good/bad, like/dislike, positive/negative, unfavorable/favorable) drawn from Holbrook and Batra (1987) as the criterion (7-point scales were applied throughout). One hundred consumers (age: M=31 years; 52% female) were exposed to print ads of a car brand and asked to complete the HED and UT scales, as well as the brand liking scale. Factor analysis confirmed the unidimensionality of the three MI scales, and their internal consistencies were highly satisfactory (α_{HED}=.93, α_{UT}=.89, and $\alpha_{BLiking}$=.94). We followed the same procedure as in Study 1 to compare the predictive validity of SI and MI measures of HED and UT, using brand liking as the criterion construct. The statistical power of our analysis was close to 1 (Cohen 1988), thus supporting the adequacy of the sample size. Table 2 summarizes the results.

[1] We also replicated Study 1 on a separate sample of 108 students from a major US university and found consistent results. The detailed results of this study are available from the authors upon request.

Table 1 Predictive validity of single and multi-item measures of A_{Ad} and A_{Brand} (Study 1)

	r[a]	*p*[b]	Bootstrap means (200 samples)	# MI better[c]	# SI better[d]	r[a]	*p*[b]	Bootstrap means (200 samples)	# MI better[c]	# SI better[d]
A_{Ad}	A: Insurance (*n*=273)					C: Jeans (*n*=289)				
Multi-item scale	.82	–	.81	–	–	.81	–	.81	–	–
Like/dislike	.78	.01	.78	136	0	.76	.00	.75	182	0
Good/bad	.78	.01	.78	155	0	.75	.00	.75	170	0
Pleasant/unpleasant	.67	.00	.67	200	0	.71	.00	.70	196	0
Informative/uninformative	.48	.00	.49	200	0	.40	.00	.39	200	0
Favorable/unfavorable	.71	.00	.71	200	0	.66	.00	.65	200	0
	B: Pain relievers (*n*=277)					D: Coffee (*n*=252)				
Multi-item scale	.81	–	.81	–	–	.82	–	.81	–	–
Like/dislike	.79	.09	.79	69	0	.73	.00	.72	198	0
Good/bad	.73	.00	.72	199	0	.74	.00	.73	193	0
Pleasant/unpleasant	.62	.00	.62	200	0	.70	.00	.69	200	0
Informative/uninformative	.38	.00	.38	200	0	.53	.00	.53	200	0
Favorable/unfavorable	.73	.00	.73	194	0	.66	.00	.66	200	0
A_{Brand}	E: Insurance (*n*=273)					G: Jeans (*n*=289)				
Multi-item scale	.64	–	.65	–	–	.72	–	.72	–	–
Good/bad	.60	.02	.60	142	0	.63	.00	.63	200	6
Like/dislike	.61	.04	.61	97	0	.65	.00	.65	200	0
Pleasant/unpleasant	.60	.01	.59	161	0	.69	.03	.69	124	0
Useful/useless	.47	.00	.46	200	0	.57	.00	.56	200	0
	F: Pain relievers (*n*=277)					H: Coffee (*n*=252)				
Multi-item scale	.65	–	.65	–	–	.61	–	.61	–	–
Good/bad	.63	.24	.63	41	2	.57	.05	.57	100	0
Like/dislike	.59	.00	.59	188	0	.56	.01	.56	146	0
Pleasant/unpleasant	.65	.49	.65	8	14	.53	.00	.53	192	0
Useful/useless	.37	.00	.37	200	0	.44	.00	.44	200	0

[a] correlation coefficient

[b] p-value of Meng et al.'s (1992) test

[c] number of bootstrap samples in which the MI scale performs significantly better than the SI predictor

[d] number of bootstrap samples in which the SI predictor performs significantly better than the MI scale

Table 2 Predictive validity of single- and multi-item measures of hedonism and utilitarianism (Study 2)

	r[a]	*p*[b]	Bootstrap means (200 samples)	# MI better[c]	# SI better[d]
A: Hedonism					
Multi-item scale	.63	–	.63	–	–
Not enjoyable/enjoyable	.63	.49	.63	13	13
Dull/exciting	.55	.02	.55	112	0
Not thrilling/thrilling	.49	.00	.48	188	0
Not fun/fun	.56	.01	.55	162	0
Not delightful/delightful	.55	.02	.53	163	0
B: Utilitarianism					
Multi-item scale	.39	–	.38	–	–
Not functional/functional	.32	.14	.31	50	1
Effective/ineffective	.33	.13	.32	48	1
Helpful/unhelpful	.32	.07	.30	89	0
Useful/useless	.39	.49	.38	10	9
Practical/impractical	.26	.03	.26	116	0

[a] correlation coefficient

[b] p-value of Meng et al.'s (1992) test

[c] number of bootstrap samples in which the MI scale performs significantly better than the SI predictor

[d] number of bootstrap samples in which the SI predictor performs significantly better than the MI scale

Focusing on HED (Table 2A), it can be seen that only not enjoyable/enjoyable has a predictive validity comparable to that of the MI scale; the performance of the other four items is consistently worse. This is supported by the bootstrapping results which show that, with the exception of not enjoyable/enjoyable, the MI scale significantly outperforms the other items in the vast majority of sample runs. A rather different picture emerges when looking at the results relating to UT (Table 2B). Here, most of the individual items display comparable predictive validity as the MI scale and this is confirmed by the bootstrap analyses.

The results of Study 2 thus suggest that while a researcher would have a good chance of selecting a "good" SI to replace the full UT scale (as four of the five items have comparable predictive validity as the MI scale), the odds are much worse for the HED scale (where only one of the five items matches the full scale in terms of predictive validity). This indicates that the relative performance of SI measures is construct-specific. Whether the predictive validity of SI measures is also stimulus-specific is addressed in Study 3.

Study 3

In Study 3, we used the same constructs and MI scales as in Study 2 but asked a new sample of 50 consumers (age: *M*= 30 years, 40% female) to rate four brands of cars in a within-subjects design. As in Study 2, we compared the predictive validity of MI scales of HED (α=.94) and UT (α=.94) attitude dimensions with that of the individual items comprising each scale; again, the four-item brand liking scale (α=.96) was used as the criterion. Table 3 summarizes the results. Statistical power was again very high (>.95), speaking against sample size concerns (Cohen 1988).

Focusing initially on HED, it can be seen that only not thrilling/thrilling consistently displays similar predictive validity as its MI counterpart across all four stimuli brands (see Table 3A, C, E, and G). The performance of the other items is unstable; sometimes their predictive validity matches that of the full scale (e.g., see not enjoyable/enjoyable for brands B and C) but sometimes not (e.g., see not enjoyable/enjoyable for brands A and D). The results relating to UT paint a similar picture; there is hardly an item that consistently matches the predictive validity of the MI measure across all four brands (see Tables 3B, D, F, and H). Moreover, a replication of Study 3 on a fresh sample of 50 consumers (age: *M*=32 years; 59% female; α_{HED}=.96, α_{UT}=.94, $\alpha_{BLiking}$=.97) but using brands of wristwatches rather than cars as stimuli produced similar findings: the predictive performance of SI compared to that of MI scales was not consistent across brands (this applies to both HED and UT as predictors of brand liking). Thus, SI performance appears to be not only construct specific (see Study 2 earlier) but also stimulus specific.

In summary, Studies 1–3 demonstrate that while SI measures *can*, in specific applications, produce a comparable level of predictive validity as MI scales, there is no guarantee that they *will*; as Stanton et al. (2002, p. 172) point out, "because correlations vary across samples, instruments and administration contexts, an item that predicts an external criterion best in one sample may not do so in another." The key question now becomes whether it is possible to identify *specific conditions* under which the predictive performance of SI measures is likely to equal (or even exceed) that of MI scales. We address this question by undertaking a Monte Carlo simulation study and use its findings to provide guidelines regarding when to use SI measures.

Table 3 Predictive validity of single- and multi-item measures of hedonism and utilitarianism (Study 3)

	r^a	p^b	Bootstrap means (200 samples)	# MI better[c]	# SI better[d]	r^a	p^b	Bootstrap means (200 samples)	# MI better[c]	# SI better[d]
	A: Brand A (Hedonism)					C: Brand B (Hedonism)				
Multi-item scale	.65	–	.65	–	–	.63	–	.61	–	–
Not enjoyable/enjoyable	.42	.00	.41	199	0	.57	.16	.55	53	1
Dull/exciting	.64	.45	.64	5	3	.52	.00	.51	173	0
Not thrilling/thrilling	.62	.26	.62	25	3	.56	.11	.54	66	0
Not fun/fun	.65	.49	.64	31	19	.60	.18	.58	45	0
Not delightful/delightful	.61	.16	.61	54	0	.63	.42	.62	10	7
	B: Brand A (Utilitarianism)					D: Brand B (Utilitarianism)				
Multi-item scale	.48	–	.48	–	–	.41	–	.39	–	–
Not functional/functional	.40	.08	.40	83	4	.31	.05	.30	99	2
Effective/ineffective	.45	.24	.44	47	4	.37	.28	.35	28	3
Helpful/unhelpful	.39	.05	.38	112	0	.31	.06	.30	96	1
Useful/useless	.45	.27	.45	51	11	.37	.24	.35	19	1
Practical/impractical	.51	.28	.52	0	33	.44	.32	.43	6	24
	E: Brand C (Hedonism)					G: Brand D (Hedonism)				
Multi-item scale	.53	–	.54	–	–	.53	–	.54	–	–
Not enjoyable/enjoyable	.48	.20	.47	51	3	.33	.00	.33	168	0
Dull/exciting	.52	.40	.52	32	10	.49	.23	.48	55	2
Not thrilling/thrilling	.53	.47	.53	22	10	.51	.34	.51	31	5
Not fun/fun	.47	.07	.47	85	1	.41	.02	.43	114	1
Not delightful/delightful	.42	.01	.42	154	0	.62	.05	.62	0	99
	F: Brand C (Utilitarianism)					H: Brand D (Utilitarianism)				
Multi-item scale	.63	–	.63	–	–	.49	–	.49	–	–
Not functional/functional	.62	.41	.61	35	8	.46	.27	.45	40	4
Effective/ineffective	.52	.02	.52	134	0	.44	.20	.44	35	0
Helpful/unhelpful	.56	.04	.56	113	0	.48	.43	.48	21	8
Useful/useless	.57	.09	.56	71	0	.44	.19	.43	61	5
Practical/impractical	.60	.31	.60	22	3	.37	.06	.37	94	0

[a] correlation coefficient

[b] p-value of Meng et al.'s (1992) test

[c] number of bootstrap samples in which the MI scale performs significantly better than the SI predictor

[d] number of bootstrap samples in which the SI predictor performs significantly better than the MI scale

Monte Carlo simulation experiment

Simulation design

Our simulation design is based on a two-construct model where one (predictor) construct predicts the other (criterion) construct (see Bergkvist and Rossiter 2007, 2009). Specifically, we compare the performance of MI scales with that of SI measures when predicting an MI criterion. We manipulate seven data and measurement characteristics (hereafter referred to as "factors") as described below:

Factor 1: Inter-item correlations among items of the predictor construct: .20 to .90 in steps of .10.
Factor 2: Inter-item correlations among items of the criterion construct: .20 to .90 in steps of .10.
Factor 3: Number of items in the predictor construct: 3, 5, and 7.
Factor 4: Number of items in the criterion construct: 3, 5, and 7.
Factor 5: Sample size: 50, 100, and 400.
Factor 6: Cross-item correlations between the predictor construct's items and the criterion construct's items: .20 to .80 in steps of .10.
Factor 7: Differences in cross-item correlations between the designated SI predictor and the *remaining* predictor items with the criterion construct's items: −.20 to .20 in steps of .10.

The choice of factors and their levels draws on Churchill and Peter (1984) as well as Peter and Churchill's (1986) studies which evaluate the effects of measurement and sample characteristics on scale reliability and, finally, on different types of validity. Their meta-analytic findings show that a construct's correlations with other constructs are positively influenced by the predictor construct's reliability as a result of an increase in systematic variance. The more homogenous the items are, the fewer of them are needed to measure a construct, and the more likely it is that SI and MI scales match up. In the extreme, if all items are identical, a sample of one (i.e., an SI) would be adequate to represent the item-set (Fuchs and Diamantopoulos 2009). It is this line of argument that prompts Loo (2002, p. 73) to recommend that "single-item measures can be considered only if the single item reflects a homogenous construct, as indicated by a high internal consistency reliability coefficient." Likewise, Gorsuch and McFarland (1972, p. 60) state that "the increase in precision gained from using multi-item scales instead of single-item scales generally arises from the increase in precision due to higher reliability."

Reliability (in an internal consistency sense) thus appears to be a potentially important influence impacting the relative performance of SI versus MI measures. Given that estimates of internal consistency such as composite reliability and Cronbach's alpha are partly a function of the strength of relationships among the items of a scale (Nunnally and Bernstein 1994), we manipulated the inter-item correlations in the measurement models of the predictor (Factor 1) and the criterion construct (Factor 2). To cover a broad range of data constellations, we varied the correlations in steps of .10 on a range of .20 to .90.

Reliability is also a function of the number of items in the scale (Nunnally and Bernstein 1994). When combining several items into a scale, random error that is inherent in every item is averaged out, which leads to increased levels of reliability (Ryan et al. 1995). More items also provide a better approximation of continuous data as they allow for a greater variation in response patterns (Bergkvist and Rossiter 2007; Churchill 1979). This increased variability is the reason why MI scales should exhibit higher levels of predictive validity with other variables when compared to SI scales (Bergkvist and Rossiter 2007; Kwon and Trail 2005). In addition, MI scales offset categorization errors and tend to be more normal and less skewed, which generally increases maximum correlation with a criterion variable (Churchill 1979; Nunnally and Bernstein 1994; Ryan et al. 1995). Consequently, we varied the number of items (3, 5, and 7 items) of both the MI predictor (Factor 3) and the MI criterion constructs (Factor 4) using a format of five response categories.[2]

Another potentially important factor is sample size, because increased precision in estimation can also be gained by using more subjects (e.g., Churchill and Peter 1984; Peter and Churchill 1986). Sample size may also influence predictive validity via the mediating factor reliability. However, while some studies report a negative relationship between sample size and alpha (e.g., Kopalle and Lehmann 1997), other studies found that alpha is insensitive to sample size (e.g., Iacobucci and Duhachek 2003). To investigate the effect of sample size (Factor 5), we varied the latter across three levels (50, 100, 400 subjects).

Comparing the performance of SI and MI measures requires assumptions regarding the true correlation structure in the model. To this end, we first established a predictor measure comprised of multiple items and specified each item's correlation with each of the criterion construct's items (ρ_1). By varying the cross-item correlations between the predictor construct's items and the criterion construct's items (Factor 6), we account for different strengths of relationships. We varied these correlations in steps of .10 on a range of .20 to .80. Low levels were considered because with weaker relationships between the items, more unspecific variance is inherent in the model (Sawyer and Ball 1981), which may potentially impact the relative performance of SI versus MI measures.

[2] The additional variation of response categories does not render any different results as the simulation examines construct correlations whose pre-specification is independent from the number of response categories.

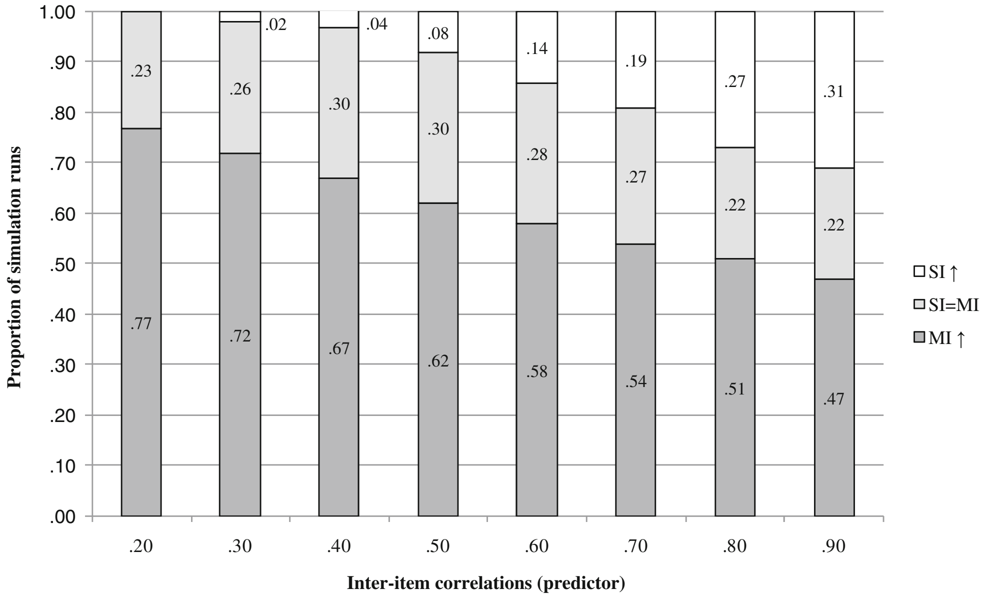


Fig. 1 Multi- vs. single-item predictive validity performance for different levels of predictor inter-item correlations

SI↑ SI predictor performs significantly better than the MI scale
MI↑ MI scale performs significantly better than the SI predictor
SI=MI SI predictor performs equally well as the MI scale

Next, we added a further item to the predictor construct (which represents the designated SI in the subsequent analyses) and pre-specified its correlation with the criterion construct's items (ρ_2), just as we did for Factor 6 above. However, we pre-specified this correlation *in relation to* ρ_1, to generate pre-specified *differences* in correlations (ρ_{diff}=ρ_1-ρ_2). Thus, in Factor 7, we varied the difference between the correlations between the predictor's multi/single item (s) and the criterion construct's (multiple) items across five levels of -.20, −.10, 0, .10, and .20. For example, a ρ_{diff} value of −.20 indicates that predictive validity is higher when the predictor is measured with an SI. Here, the chosen SI predicts the relevant criterion better than it predicts the *remaining* items in the predictor's item-set; we can thus evaluate whether an SI that correlates highly with a criterion will significantly increase the performance of the overall scale when combined with other items that only weakly contribute to the explanation of the criterion construct. Conversely, a ρ_{diff} value of .20 indicates that the chosen SI's predictive validity lags behind that of the remaining items. In this case, the SI only contributes weakly to the explanation of the criterion construct; thus, we also explicitly account for situations in which the researcher chooses the "wrong" SI from a set of potential candidate items.

In line with prior simulation studies and given the large number of factor level combinations, we used five replications for each combination. Ex post analyses of the results showed that these were robust across replications. As not all potential factor combinations are feasible,[3] we only considered situations that can actually occur in practice (a total of 181,758 datasets).

Data generation procedure and model estimation

For data generation, we adopted Kaiser et al.'s (2011) procedure to draw categorized random numbers from a given correlation matrix of predictor and criterion constructs' items (see Appendix). Following Bergkvist and Rossiter (2007, 2009), we computed MI scale scores for the predictor and criterion constructs by summing individual item scores. Predictive validity was subsequently assessed by comparing the correlations between the predictor and criterion constructs using Meng et al.'s (1992) test, leading to one of three outcomes: (1) the predictive validity of the MI scale is significantly higher (at p<.05 or better) than that of the SI predictor, (2) there is no significant difference between the predictive validity of the MI scale and that of the SI predictor, and (3) the predictive validity of the SI predictor is significantly higher (at p<.05 or better) than that of the MI scale.

Simulation results

The MI scale significantly outperforms the SI predictor in 59.90% (108,867 runs) of all simulation runs. The SI predictor performs significantly better in only 14.10% (25,613 runs) of all runs, whereas in the remaining cases (26.00%, 47,278 runs), there is no significant difference between the two specifications. A χ^2 goodness-of-fit test on the observed frequencies reveals that the number of occasions where the MI scale (significantly) outperforms the SI predictor is

[3] For example, it is impossible to have low values in ρ_1 and low inter-item correlations in the predictor construct while at the same time having high inter-item correlations among items of the criterion construct.

Table 4 Multi- vs. single-item predictive validity performance for different numbers of items in the predictor and criterion constructs

	Criterion construct									
		SI ↑	SI=MI	MI ↑	SI ↑	SI=MI	MI ↑	SI ↑	SI=MI	MI ↑
	# of items	3 items			5 items			7 items		
Predictor construct	3 items	.13	.29	.58	.14	.26	.60	.15	.25	.60
	5 items	.14	.28	.59	.14	.25	.61	.15	.25	.60
	7 items	.13	.27	.59	.14	.25	.61	.15	.24	.61

SI↑ SI predictor performs significantly better than the MI scale

MI↑ MI scale performs significantly better than the SI predictor

SI=MI SI predictor performs equally well as the MI scale

Table entries are proportions of simulation runs

significantly higher than the number of occasions where no difference is observed or when the reverse is the case ($\chi^2(1)$= 7,120.46, $p<.0001$). Moreover, the average R^2 differences are significantly higher (.12 units) when the MI scale is a better predictor ($t_{134,479}$=313.865, $p<.001$), indicating that explanatory power is also greater when the MI scale performs better than the SI. A discussion of the detailed results follows.

Inter-item correlations As Fig. 1 shows, for most levels of inter-item correlations (Factor 1), the MI predictor clearly outperforms the SI predictor; this effect, however, decreases with increasing magnitude of inter-item correlations. Practically identical results (not displayed in Fig. 1) were obtained for the inter-item correlations of the criterion construct (Factor 2).

Number of items Neither the number of items in the predictor construct (Factor 3) nor that in the criterion construct (Factor 4) has an influence on SI performance (Table 4). Regardless of the number of items in these constructs, the MI scale significantly outperforms single items in about 60% of the cases; the proportion of cases where the SI predictor performs significantly better than the MI scale is also constant (around 14%). Thus, choosing an item from an existing short scale versus a longer scale does not change the probability that the chosen item's predictive validity performance will be closer to the performance of the full scale.

Sample size For a small sample size of N=50 (Factor 5), the MI scale significantly outperforms the SI in 51% of the cases, whereas for N=400, this number increases to 70% of the cases (Table 5). Lower sample sizes lead to proportionally more simulation runs in which SI and MI scales perform equally well.

Cronbach's alpha To capture the joint effects of different levels of inter-item correlations (Factor 1), different numbers of items in the predictor construct (Factor 3), and sample size (Factor 5), we computed Cronbach's alpha values of the predictor construct for each simulation run (Fig. 2).[4] Results parallel the findings on Factor 1 and underline the MI scale's superior performance across most alpha levels. SI performance is equivalent to the performance of the MI scale only for very high alpha levels (>.90).

Cross-item correlations between the predictor construct's items and the criterion construct's items Figure 3 shows that the predictive validity of MI vis-à-vis SI scales increases with increased levels of cross-item correlations (Factor 6). High cross-item correlations leave less unique variance to be potentially explained by single items and this is mirrored in the MI scale's superior performance in these situations.

Differences in cross-item correlations Table 6 shows that SI measures perform favorably only in situations in which the chosen SI exhibits much higher correlations with the criterion ($\rho_{\text{diff}}=-.20$) than the MI predictor (Factor 7). In situations where the chosen SI and the remaining items are similarly related to the criterion constructs (i.e., when $\rho_{\text{diff}}=0$), the MI scale's predictive validity is superior in two-thirds of runs. When the chosen SI is a "poor" item (i.e., for $\rho_{\text{diff}}=.10$ and .20), predictive validity is likely to be lower in more than 90% of the cases.

Meta-analyses of the simulation results

We meta-analyzed the simulation results by fitting a logistic regression model, in which the dependent variable was coded as follows: 0=the SI predictor performs significantly better than or as well as the MI scale (SI≥MI), and 1=the MI scale performs significantly better than the SI predictor (MI>SI).

[4] Note, that alpha values *result* from factor level combinations; we did *not* explicitly include alpha as a distinct design factor in our simulation. Thus, also very low values (considered unreasonable in empirical research) are included in Fig. 2.

Table 5 Multi- vs. single-item predictive validity performance for different sample sizes

	Sample size		
	50	100	400
SI ↑	.11	.14	.17
SI=MI	.38	.27	.13
MI ↑	.51	.59	.70

SI↑ SI predictor performs significantly better than the MI scale

MI↑ MI scale performs significantly better than the SI predictor

SI=MI SI predictor performs equally well as the MI scale

Table entries are proportions of simulation runs.

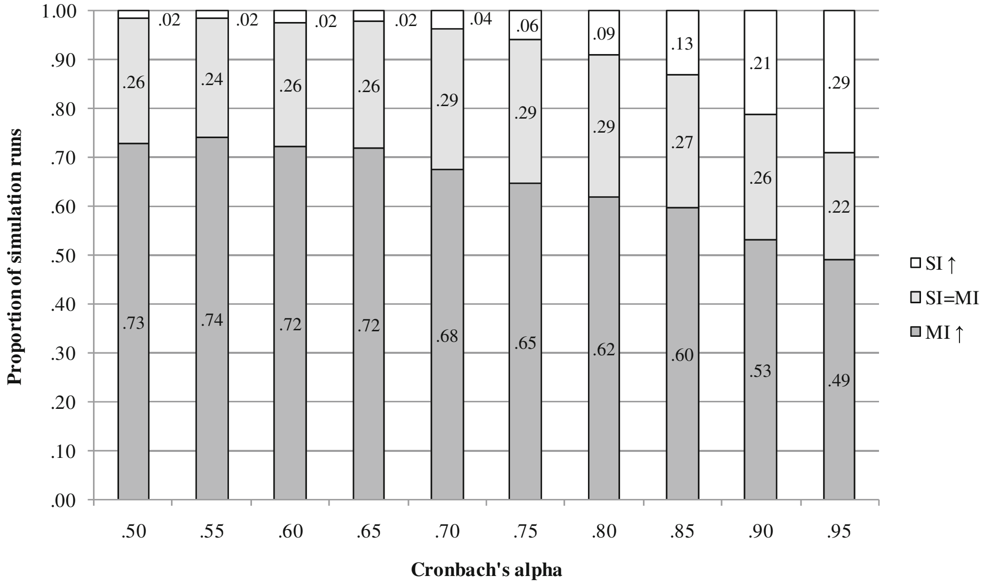
All goodness-of-fit statistics indicate a highly satisfactory model fit (Table 7). The parameter estimates clearly show that the simulation design characteristics investigated (i.e., Factors 1–7) exert a significant influence on the performance of MI versus SI scales. Specifically, the log odds of the MI scale performing better than the SI increase with lower inter-item correlations, increasing numbers of items, increasing sample size, higher cross-item correlations, and increasing differences in cross-item correlations (ρ_{diff}). We further assessed two-way interaction effects for the most important factors. The results show that the log odds of the MI scale performing better than the SI measure decrease with higher inter-item correlations in the predictor and criterion constructs and increasing cross-item correlations (Factor 1×Factor 6, and Factor 2×Factor 6). Interactions between cross-item correlations (Factor 6) and differences in cross-item correlations (Factor 7) reveal a consistently positive influence on the log odds.

Guidelines for researchers

Figure 4 shows specific conditions favoring the use of SI measures that researchers can use in the early stages of research design to decide whether or not to adopt single items. For existing constructs, researchers can readily consult scale handbooks (e.g., Bearden et al. 2011; Bruner et al. 2005) to obtain prior estimates of reliability and previous published applications of the scale to derive expectations regarding the magnitude of relationships with outcome variables (i.e., small, moderate, large). In the case of scales for new or adapted constructs, pretesting can be used to obtain rough estimates of reliability, effect sizes, etc.

According to the simulation results, for sample sizes smaller than 50, relying on an SI is likely to generate comparable predictive validity as that of the full scale. Thus, the use of SI measures could be considered if sample size is restricted due to, for example, budget constraints, difficulties in recruiting respondents, limited population size or the need to collect dyadic data (Fuchs and Diamantopoulos 2009). Similarly, single items appear to be a reasonably safe bet when rather weak effects are expected (i.e., cross-item correlations less than .30). When inter-item correlations are above .80 or the construct's alpha values higher than .90—that is, the measure is highly homogenous in an internal

Fig. 2 Multi- vs. single-item predictive validity performance for different levels of Cronbach's alpha

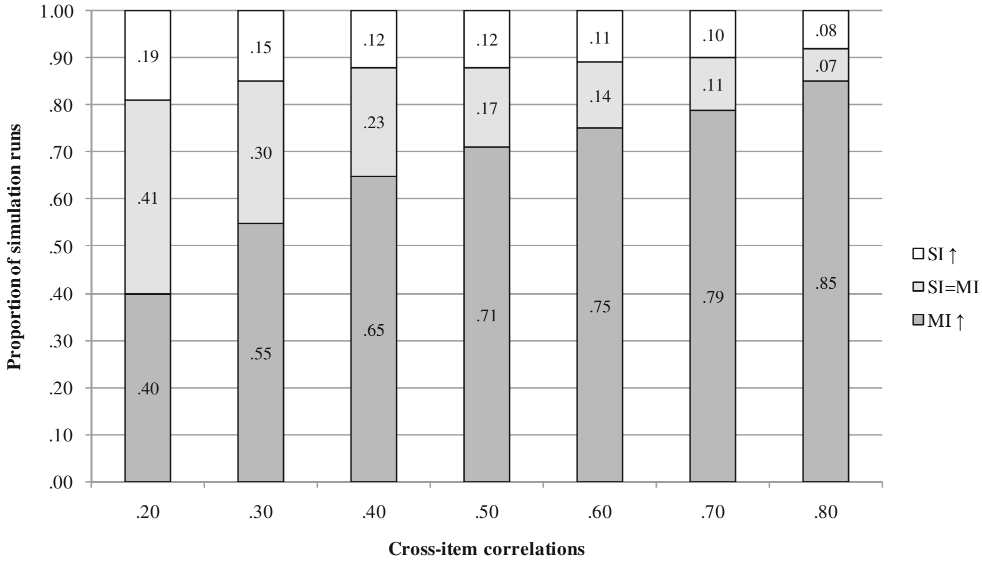


SI↑ SI predictor performs significantly better than the MI scale

MI↑ MI scale performs significantly better than the SI predictor

SI=MI SI predictor performs equally well as the MI scale

Fig. 3 Multi- vs. single-item performance for different levels of cross-item correlations



SI↑ SI predictor performs significantly better than the MI scale

MI↑ MI scale performs significantly better than the SI predictor

SI=MI SI predictor performs equally well as the MI scale

consistency sense—the use of single items should also be considered. Note that these thresholds are way above those recommended in the measurement literature (e.g., see DeVellis 2003; Nunnally and Bernstein 1994; Robinson et al. 1991) and, if attained, likely to be indicative of semantic redundancy which adversely effects content validity (Boyle 1991; Rossiter 2002). Indeed, when items are highly redundant, “information content can be negatively affected by scales with multiple items” (Drolet and Morrison 2001, p. 197) and, thus, the use of SIs is advisable also from a conceptual perspective. Having said that, high inter-item correlations should not *automatically* be equated with semantically redundant items, because items measuring different aspects of a construct may also be highly related. Mechanically replacing an MI scale with an SI when inter-item correlations and alpha values indicate a highly homogenous construct can have adverse consequences for the measure’s content validity (Grapentine 2001). Therefore, researchers should carefully scrutinize the full scale (either based on prior research results or those from a pretest in case of newly developed measures) and determine the extent to which items are semantically redundant and whether all facets of the construct domain are captured. This judgment should be conducted by at least two expert coders independently to ensure a high degree of objectivity. Needless to say, in order to enable such an assessment, researchers should not only report alpha values for their MI scales but also the relevant inter-item correlations.

Table 6 Multi- vs. single-item predictive validity performance for differences in cross-item correlations

	Differences in cross-item correlations				
	−.20	−.10	0	.10	.20
SI ↑	.54	.26	.00	.00	.01
SI=MI	.37	.48	.34	.08	.06
MI ↑	.09	.26	.66	.92	.93

SI↑ SI predictor performs significantly better than the MI scale

MI↑ MI scale performs significantly better than the SI predictor

SI=MI SI predictor performs equally well as the MI scale

Table entries are proportions of simulation runs

Overall, SI measures seem to be a viable option in exploratory research situations where typically weaker effect sizes are expected and smaller samples are used (Stebbins 2001). In exploratory studies, the research objective is more to map out the main effects in a nomological network rather than to identify detailed aspects of constructs and their interrelations (Lee et al. 2000). Under such conditions, it is often sufficient to ask a single “global” question which allows a respondent to “consider all aspects and individual preferences of the certain aspects of the construct being measured” (Nagy 2002, p. 79). With “global” SI 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). However, constructs must be accurately described and made clear to respondents when measured with an SI measure (Sackett and Larson 1990; Fuchs and Diamantopoulos 2009). This is because SI measures require more abstract thinking as opposed to MI scales (Sloan et al. 2002) and therefore may be too vague for respondents to be “correctly” answered. Thus, the use of SI measures in exploratory research settings is

Table 7 Logistic regression results

		Main effects[a]		Interaction effects[b]	
Factor		B	z	B	z
	Intercept	3.72***	−82.86	−9.62***	−71.77
1	Inter-item correlations (predictor)	−6.30***	−134.00	3.48***	30.15
2	Inter-item correlations (criterion)	−1.86***	−48.35	−1.57***	−15.35
3	#items criterion=5	.41***	21.21	.50***	24.08
	#items criterion=7	.60***	30.28	.71***	34.23
4	# items predictor=5	.14***	7.17	.15***	7.36
	# items predictor=7	.20***	10.03	.22***	10.59
5	Sample size=100	.95***	442.02	1.05***	50.34
	Sample size=400	2.13***	98.29	2.36***	101.09
6	Cross-item correlations	9.98***	139.52	26.57***	75.88
7	ρ_{diff}=-.10	1.51***	54.54	1.08***	10.51
	ρ_{diff}=0	4.06***	136.72	2.16***	21.66
	ρ_{diff}=.10	6.89***	176.42	4.31***	39.10
	ρ_{diff}=.20	7.14***	172.20	2.94***	21.03
1×6	Inter-item correlations (predictor)×Cross-item correlations	–	–	−30.76***	−93.59
2×6	Inter-item correlations (criterion)×Cross-item correlations	–	–	−1.36***	−3.94
6×7	Cross-item correlations×ρ_{diff}=-.10	–	–	2.11***	8.54
	Cross-item correlations×ρ_{diff}=0	–	–	7.14***	28.07
	Cross-item correlations×ρ_{diff}=.10	–	–	9.39***	30.71
	Cross-item correlations×ρ_{diff}=.20	–	–	16.35***	32.77

[a] Nagelkerke's R^2=.74, c-index=.95, Gamma=.90

[b] Nagelkerke's R^2=.77, c-index=.96, Gamma=.92

*** p<.001, ** p<.01, * p<.05

Reference categories: #items predictor=3, #items criterion=3, sample size=50, ρ_{diff}=−.20

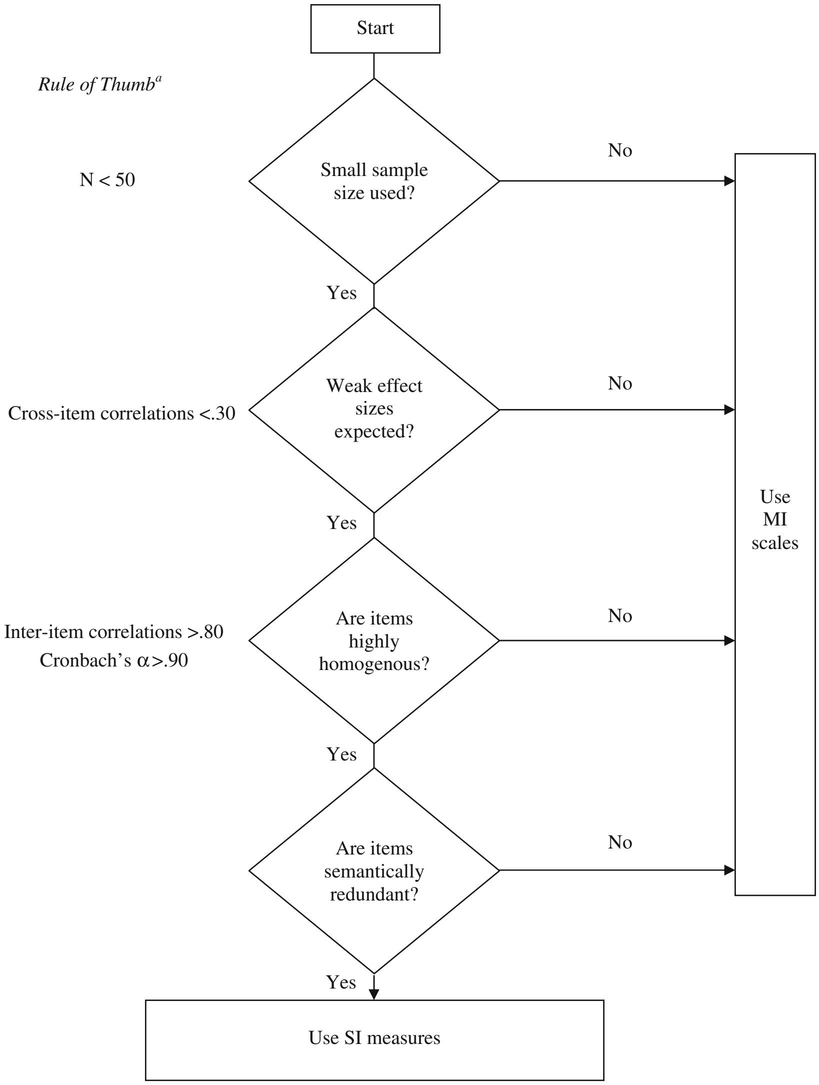
advisable when the construct is very concrete and it can be reasonably assumed that there is virtually unanimous agreement among respondents as to what characteristic is being measured (Rossiter 2002).

Conclusions

Some twenty years ago, Sackett and Larson (1990) argued that when a construct is narrow in scope, unidimensional, and unambiguous to the respondent, using an SI is the best measurement approach. More recently, Rossiter (2002, p. 313) echoed this view by stating that "when an attribute is judged to be concrete, there is no need to use more than a single item […] to measure it in the sale" and subsequently offered evidence indicating similar predictive validity of an SI predictor as compared to an MI scale (Bergkvist and Rossiter 2007, 2009).

Despite such initial evidence, however, abandoning established MI scales (even for concrete constructs) in favor of SI measures is not without risk for researchers. As shown by Studies 1–3, SI predictive validity performance can be quite variable across constructs, product categories, and stimuli (i.e., brands). Thus, an SI which may perform as well as the MI scale in one context may not do so in another. In an attempt to explain the observed variability we sought to identify specific conditions under which the predictive validity performance of an SI is likely to at least match that of an MI scale. We did this by means of a simulation study in which several important factors identified in literature were manipulated and their effects on the relative performance of SI versus MI measures investigated. Our results showed that opting for SI measures in most empirical settings is a risky decision as the set of circumstances that would favor their use is unlikely to be frequently encountered in practice. While there *are* circumstances in which single items could be legitimately employed (see Fig. 4), in most instances, researchers would be well-advised to follow "conventional wisdom" and employ MI scales in empirical investigations, as recommended by extant literature (e.g., Churchill 1979; DeVellis 2003; Netemeyer et al. 2003; Spector 1992). Even opting for a "halfway" solution and using two (or three) items instead of the full scale is not advisable. While such an approach is intuitively appealing as it appears to offer a "margin of safety," it is not without problems. First, often only one item in an MI scale has comparable predictive validity with the full scale (e.g., see not enjoyable/enjoyable in Table 2A). Combining one "good" item with a "poor" item (e.g., see not thrilling/thrilling in Table 2A) seems hardly a sensible strategy to follow. Second, even if two items can be identified with comparable performances as the MI scale (e.g., see not functional/functional and effective/ineffective in Table 2B), the resulting measurement

Fig. 4 Guidelines for choosing between MI and SI measures


Start
Rule of Thumb[a]
N < 50
Small sample size used?
No
Yes
Cross-item correlations <.30
Weak effect sizes expected?
No
Yes
Inter-item correlations >.80
Cronbach's α>.90
Are items highly homogenous?
No
Yes
Are items semantically redundant?
No
Yes
Use MI scales
Use SI measures

[a] based on simulation results

model is underidentified and cannot be estimated (let alone tested).[5] A three-indicator model, on the other hand, can be estimated but will always return "perfect" fit as the model is saturated.[6] While introduction of equality constraints on the indicator loadings and/or error variances can be used to overcome these problems, such modifications are not consistent with congeneric measurement (Darden et al. 1984; Jöreskog 1971). Thus a minimum of four items is generally recommended to enable independent estimation and testing of measurement models (e.g., Bollen 1989; Long 1983). Needless to say, if there are four items, one is de facto employing an MI scale and, therefore, the only alternative really worth considering is an SI (which brings us full circle).

Limitations and future research

While our study makes a first step in investigating the conditions under which SI measures are likely to have the same predictive ability as MI scales, several issues require further study. First, future studies should compare SI and MI measures in more complex model set-ups. For example, while our simulation design considered uniform levels of predictor inter-item correlations, it would be interesting to vary the levels of correlations between the designated SI and the remaining items in the scale. Second, prior literature has argued that the selection of an SI versus MI measure should take into account the role of the focal construct in the research design (Fuchs and Diamantopoulos 2009). Boyd

[5] Four parameters (two loadings and two error variances) need to be estimated, but there are only three non-redundant elements in the covariance matrix (the two variances of the indicators and the covariance between them).

[6] With a saturated model, there are zero degrees of freedom as all available information is used to generate a unique solution for the parameter estimates; there is no remaining information to enable testing of the model.

et al. (2005), for example, show that researchers routinely use single items as control or moderator variables in empirical studies. Thus, evaluating SI performance when measuring a moderator variable is a promising area for future research. Third, an investigation of the stability of SI performance in longitudinal studies would provide additional insights on the extent to which SI measures offer a realistic alternative to that of MI scales in practical applications. Fourth, examining the extent to which granularity in response options impacts the predictive validity performance of SI measures is another issue worthy of future study. Last but not least, empirically-based guidelines for *selecting* a single item are sorely needed for those instances favoring the use of SI measures. Given that, in practical applications one needs to select a "good" item *before* data collection, and given that one would not normally collect data on the *other* items comprising the scale, it is important to know ex ante whether the selected SI represents a good choice or not. A comparison of the relative effectiveness of alternative item selection methods therefore constitutes an important avenue for future research; the literature on scale reduction (e.g., Moore et al. 2002; Smith and McCarthy 1995; Stanton et al. 2002) may provide a useful starting point for such research endeavors.

Appendix: data generation procedure

1 Based on the pre-specified correlation structure of the categorized variables cor(A) (A=A$_1$,…A$_k$), a new correlation structure cor(X) (X=X$_1$,…,X$_k$) is established by transforming the categorized correlations cor(A$_i$, A$_j$) into the corresponding normal correlations cor(X$_i$, X$_j$). This is done pairwise for all combinations of A$_i$, A$_j$, i≠j.

2 A sample is drawn from the k-dimensional normal distribution with the mean vector 0 and covariance matrix cor(X) from step 1 to derive a sample matrix S. This random drawing ensures the inclusion of variation in the data.

3 Subsequently, this sample matrix S is transformed into a categorized matrix using

$$\widetilde{S}_{[nm]} = s : \left\{F_{A_m}(s-1) < \Phi\left(S_{[nm]}\right) < F_{A_m}(s)\right\},$$

for $s \in \{1, ..., s_{A_m}\}$ and Φ being the univariate standard normal distribution function.The result is a data matrix $\widetilde{S}_{[nm]}$ with categorized values which have a given probability for each category and correlation matrix cor(A).

References

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research, 12* (4), 453–460.

Bagozzi, R. P. (1982). The role of measurement in theory construction and hypothesis testing: Toward a holistic model. In C. Fornell (Ed.), *A second generation of multivariate analysis. Measurement and evaluation. Vol. 2* (pp. 5–23). New York: Greenwood Publishing Group.

Batra, R., & Ahtola, O. T. (1991). Measuring the hedonic and utilitarian sources of consumer attitudes. *Marketing Letters, 2*(2), 159–170.

Bearden, W. O., Netemeyer, R. G., & Haws, K. L. (2011). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research* (3rd ed.). Thousand Oaks: Sage Publications, Inc.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*(2), 175–184.

Bergkvist, L., & Rossiter, J. R. (2009). Tailor-made single-item measures of doubly concrete constructs. *International Journal of Advertising, 28*(4), 607–621.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin, 110*(2), 305–314.

Boyd, B. K., Gove, S., & Hitt, M. A. (2005). Construct measurement in strategic management research: illusion or reality? *Strategic Management Journal, 26*(3), 239–257.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12*(3), 291–294.

Bruner, G. C., Hensel, P. J., & James, K. E. (2005). *Marketing scales handbook: A compilation of multi-Item measures* (Vol. 4). Chicago: American Marketing Association.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills: Sage.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale: Lawrence Erlbaum Associates.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16* (1), 64–73.

Churchill, G. A., Jr., & Peter, J.-P. (1984). Research design effects on the reliability of rating scales: a meta-analysis. *Journal of Marketing Research, 21*(4), 360–375.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-validation for prediction. *Journal of Marketing Research, 24*(3), 271–279.

Darden, W. S., Carlson, M., & Hampton, R. D. (1984). Issues in fitting theoretical and measurement models in marketing. *Journal of Business Research, 12*(3), 273–296.

De Boer, A. G. E. M., van Lanschot, J. J. B., Stalmeier, P. F. M., van Sandick, J. W., Hulscher, J. B. F., de Haes, J. C. J. M., & Sprangers, M. A. G. (2004). Is a single-item visual analogue scale as valid, reliable, and responsive as multi-item scales in measuring quality of life? *Quality of Life Research, 13*(2), 311–320.

De Jong, M. G., Lehmann, D. R., Netzer, O. (2010). State-dependence effects in surveys: A global investigation. *Working Paper.*

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Thousand Oaks: Sage Publications, Inc.

Diamantopoulos, A. (2005). The C-OAR-SE procedure for scale development in marketing: a comment. *International Journal of Research in Marketing, 22*(1), 1–9.

Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research, 3* (3), 196–204.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika, 68*(3), 589–599.

Ferguson, G. A. (1971). *Statistical analysis in psychology and education* (3rd ed.). New York: McGraw-Hill.

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research. *Business Administration Review, 69*(2), 195–210.

Gardner, D. G., Dunham, R. B., Cummings, L. L., & Pierce, J. L. (1989). Focus of attention at work: construct definition and empirical validation. *Journal of Occupational Psychology, 62*, 61–77.

Gorsuch, R. L., & McFarland, S. G. (1972). Single vs. multiple-item scales for measuring religious values. *Journal for the Scientific Study of Religion, 11*(1), 53–64.

Grapentine, T. H. (2001). A practitioner's comment on Aimee L. Drolet and Donald G. Morrison's "Do we really need multiple-item measures in service research?". *Journal of Service Research, 4*(2), 155–158.

Holbrook, M. B., & Batra, R. (1987). Assessing the role of emotions as mediators of consumer responses to advertising. *The Journal of Consumer Research, 14*(3), 404–420.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: measuring reliability with confidence. *Journal of Consumer Psychology, 13* (4), 478–487.

Jacoby, J. (1978). Consumer research: a state of the art review. *Journal of Marketing, 42*(2), 87–96.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109–133.

Kaiser, S., Träger, D., Leisch, F. (2011). Generating correlated ordinal random values. *Technical Report No. 94*, Institute of Statistics, Ludwig-Maximilians-University Munich. Retrieved December 17, 2011from http://epub.ub.uni-muenchen.de/12157/.

Kopalle, P. K., & Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes, 70*(3), 189–197.

Kwon, H., & Trail, G. (2005). The feasibility of single-item measures in sport loyalty research. *Sport Management Review, 8*(1), 69–89.

Lee, H., Delene, L. M., Bunda, M. A., & Chankon, K. (2000). Methods of measuring health-care service quality. *Journal of Business Research, 48*(3), 233–246.

Long, J. S. (1983). *Confirmatory factor analysis*. Beverly Hills: Sage.

Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology, 17*(1), 68–75.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172–175.

Moore, K. A., Halle, T. G., Vandivere, S., & Mariner, C. L. (2002). Scaling back survey scales: how short is too short? *Sociological Methods & Research, 30*(4), 530–567.

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology, 75*, 77–86.

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

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

Okada, E. M. (2005). Justification effects on consumer choice of hedonic and utilitarian goods. *Journal of Marketing Research, 42*(1), 43–53.

Peter, J.-P. (1979). Reliability: a review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*(1), 6–17.

Peter, J.-P. (1981). Construct validity: a review of basic issues and marketing practices. *Journal of Marketing Research, 18*(2), 133–145.

Peter, J.-P., & Churchill, G. A., Jr. (1986). Relationships among research design choices and psychometric properties of rating scales: a meta analysis. *Journal of Marketing Research, 23*(1), 1–10.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 1–15). San Diego: Academic.

Rogelberg, S. G., & Stanton, J. M. (2007). Introduction understanding and dealing with organizational survey nonresponse. *Organizational Research Methods, 10*(2), 195–209.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19*(4), 305–335.

Rossiter, J. R. (2011). *Measurement for the social sciences. The C-OAR-SE method and why it must replace psychometrics*. Berlin: Springer.

Ryan, M. J., Buzas, T., & Ramaswamy, V. (1995). Making CSM a power tool—Composite indices boost the value of satisfaction measures for decision making. *Marketing Research, 7*(3), 10–16.

Sackett, P. R., & Larson, J. R. (1990). Research strategies and tactics in I/O psychology. In M. D. Dunnette, P. L. Ackerman, L. M. Hough, & H. C. Triandis (Eds.), *Handbook of industrial and organizational psychology Vol 1* (pp. 419–489). Palo Alto: Consulting Psychologists Press.

Sawyer, A. G., & Ball, D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research, 18*(3), 275–290.

Sloan, J. A., Neil, A., Cappelleri, J. C., Fairclough, D. L., Varricchio, C., & Clinical Significance Consensus Meeting Group. (2002). Assessing the clinical significance of single items relative to summated scores. *Mayo Clinic Proceedings, 77*(5), 479–487.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 300–308.

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Newbury Park: Sage Publications, Inc.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*(1), 167–193.

Stebbins, R. A. (2001). *Exploratory research in the social sciences. Sage University Papers Series on Qualitative Research Methods, Vol. 48*. Thousand Oaks: Sage Publications.

Steenkamp, J.-B. E. M., & van Trijp, H. C. M. (1991). The use of LISREL in validating marketing constructs. *International Journal of Research in Marketing, 8*(4), 283–299.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251.

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

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring hedonic and utilitarian dimensions of consumer attitude. *Journal of Marketing Research, 40*(3), 310–320.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: how good are single-item measures? *Journal of Applied Psychology, 82*(2), 247–252.