Appropriate use of single-item measures is here to stay

Lars Bergkvist

Abstract In their article, Bergkvist and Rossiter (Journal of Marketing Research, 44, 175–184, 2007) recommended marketing academics to use single-item instead of multiple-item measures for doubly concrete constructs. This recommendation was based on a study showing that the predictive validity of single-item measures was comparable to that of multiple-item measures. Kamakura (2014) presents three criticisms of Bergkvist and Rossiter’s study: (1) The correlations used to evaluate predictive validity are inflated by the presence of common-methods variance in the data, (2) the study used concurrent validity as criterion rather than predictive validity, and (3) the multiple-item measures in the study were not corrected for attenuation. A re-analysis of the data from the original study refutes the claims made by Kamakura (2014). The analysis shows that the common-methods variance in the data was negligible and that predicting delayed measures rather than concurrent measures yielded virtually identical results as in the original study. It is also shown that it is possible to estimate single-item reliabilities and correct single-item measures for attenuation, which makes them as predictively valid as multiple-item measures. Thus, there is no reason to change the conclusions and recommendations made in Bergkvist and Rossiter’s (Journal of Marketing Research, 44, 175–184, 2007) article. The present article also shows that Kamakura’s (2014) analysis of consumer panel data has limitations which casts doubts upon the conclusions drawn from the analysis results. In addition, there is a discussion of the cost, in terms of research quality, that researchers unnecessarily using multiple-items measures pay.

Keywords Single-item measures · Multiple-item measures · Psychometrics · Reliability · Validity

1 Introduction

The recommendation in Bergkvist and Rossiter (2007; hereafter B&R), based on an analysis of the predictive validity of single- and multiple-item operationalizations
of the same constructs, is that marketing researchers should use single-item measures for doubly concrete constructs. Doubly concrete constructs are constructs that have a simple, clear object (e.g., an ad or a brand) and a single and single-meaning attribute (e.g., liking) (Rossiter 2010). Examples of doubly concrete constructs include attitude toward the ad (A_Ad), brand attitude (A_Brand), and brand purchase intention (PI_Brand) (Bergkvist and Rossiter 2009). In addition, B&R recommend that marketing journals should be willing to accept papers correctly using single-item measures for publication. B&R’s analysis and recommendations were based on the fundamental insight in Rossiter (2002) that there are different types of constructs and that psychometrics, or classical test theory, are not applicable to all of them.

Kamakura (2014) has reservations about the empirical study in B&R and sees the recommendation to use single-item measures as ill advised for any type of construct. Instead, he argues that multiple-item measures “are necessary not only to improve the validity of some measurement instruments but, more importantly, to make it possible to assess and correct measurement instruments for random (non-systematic) measurement errors” (Kamakura 2014, abstract). Moreover, Kamakura presents results from an analysis of data from a consumer panel in which attitudinal measures were used to predict purchase behaviors which he argues shows that multiple-item measures outperform their single-item opposite numbers.

This article first shows that Kamakura’s (2014) criticism of B&R is ill founded and that the conclusions from the study are still valid. It then presents a critique of the consumer panel analysis in Kamakura’s article, showing that the conclusion that multiple-item measures are superior to single-item measures is not valid. This is followed by a discussion of the costs, in the form of negative effects on the quality of research, of unnecessarily using multiple-item measures.

2 Refutation of the criticisms made against B&R

There are three main criticisms of the empirical study in B&R in Kamakura (2014): (1) the correlations used to evaluate predictive validity are inflated by the presence of common-methods variance in the data, (2) B&R studied concurrent validity rather than predictive validity, and (3) it is not possible to estimate the reliability of single-item measures and, as a consequence, these measures cannot be corrected for attenuation. These criticisms will be addressed in the following sections.

2.1 Common-methods variance

According to Kamakura (2014), the results in B&R are not valid because they are biased by common-method variance in the data. Fortunately, it is possible to test the extent of common-method variance using the marker variable technique (Lindell and Whitney 2001; see also Podsakoff et al. 2003). This method uses the correlation between the focal variables and a theoretically unrelated variable to estimate the common-methods variance, which can then be used to partial out common-methods bias. In the absence of an unrelated variable, one or two of the
variables included in the research can be used as a proxy for a marker variable (Malhotra et al. 2006).

The procedure developed by Malhotra et al. (2006) prescribes that the variable with the smallest correlation with the focal variables in the study be used as marker variable. In the B&R study, the lowest correlations were with items measuring various brand benefit beliefs. However, as the benefit beliefs items were different for the four ads, they could not be used across all four ads. Instead, one of the items in a multiple-item measure of ad credibility (biased–unbiased, measured on a seven-point answer scale) was used as a marker variable across all four ads even if its correlation was consistently higher than the correlation with the benefit beliefs. For each ad, the two lowest correlations between the marker variable and the variables included in the correlational analyses in B&R were estimated. (The second-lowest correlation provides a more conservative estimate of common-methods variance.) These two correlations were then used to calculate correlation coefficients adjusted for common-methods variance using Formula 1 in Malhotra et al. (2006, p. 1868).

The marker variable analysis was applied to the six correlations that were reported for each ad in Table 1 in Kamakura (2014). The results of the analysis clearly show that common-methods variance was not a problem in the B&R dataset. For the four ads, the average difference between the unadjusted and adjusted correlations were 0.022 (painkillers), 0.006 (coffee), 0.041 (pension plan), and <0.001 (jeans). (The corresponding values using the more conservative second lowest correlations were 0.031, 0.037, 0.051, and 0.012.) Moreover, none of the correlations reported in B&R changed from significant to nonsignificant when the correlation was adjusted. Thus, the extent of common-methods variance in the data is so small that its effect on the estimated correlations is negligible.

2.2 Predictive validity

Kamakura (2014) claims that B&R did not measure predictive validity but rather concurrent validity. This claim is based on Cronbach and Meehl’s (1955, p. 282) definition of predictive validity which requires that “the criterion is obtained sometime after the test is given.” This definition of predictive validity is not universally accepted. For example, both Nunnally and Bernstein (1994) and DeVellis (2003) define predictive (or criterion-related) validity as prediction based on measurement before, during, or after the event that is predicted. In fact, both books consider concurrent validity to be one form of predictive validity.

However, irrespective of how one defines predictive validity, it could be argued that, if possible, it is better to measure the criterion variable in a separate study from the predictor. In fact, the B&R dataset was part of a larger study of advertising effectiveness measurement and the full study included measurement of $A_{Brand}$ and some other advertising effectiveness measures 4 weeks after the copy test that made up the original B&R dataset (a description of the methodology is available in Bergkvist and Rossiter

---

1 The formula for the CMV adjusted correlation is $r_{A} = \frac{r_{U} - r_{M}}{1-r_{M}}$, where $r_{A}$=the CMV adjusted correlation, $r_{M}$=the lowest correlation between the marker variable and focal variables, and $r_{U}$=the uncorrected correlation between two theoretically related variables.
Thus, the predictive validity of the single- and multiple-item measures of $A_{Ad}$ from B&R can be tested against delayed measures of $A_{Brand}$ obtained 4 weeks after the first set of measures were taken. The delayed $A_{Brand}$ measures were captured using the same set of measures as in the original B&R dataset.

The correlations between the $A_{Ad}$ measures and the delayed $A_{Brand}$ measures are shown in Table 1. There were mostly small differences between the single- and multiple-item measures: The average differences in the correlations for $A_{Ad3}$ and $A_{Ad1(G)}$ across the four ads were 0.04, 0.01, and $-0.01$ for the three dependent variables. Thus, the results with the delayed measure of $A_{Brand}$ as criterion were virtually identical to the results obtained in the original B&R study that used the immediate measure of $A_{Brand}$.

### 2.3 Reliability estimates and correction for attenuation

An often-heard argument in favor of multiple-item measures is that these, unlike single-item measures, allow for the calculation of internal consistency reliability (Churchill 1979). However, it is possible to estimate the reliability of a single-item measure. Wanous and Reichers (1996) showed how the formula for correction of attenuation (Nunnally and Bernstein 1994, p. 257) can be used to estimate the reliability of single-item measures (see also Wanous and Hudy 2001; Wanous et al. 1997). Assuming that the correlation between two error free measures of the same construct is 1.00, the correlation between the two measures and the internal consistency reliability (alpha) for the multiple-item measure, which can both be estimated from the data, can be plugged into the formula for correction of attenuation, which leaves the reliability for the single-item measure as the only unknown in the formula. Rearranging the formula makes it possible to solve for

<table>
<thead>
<tr>
<th>Predictors of delayed $A_{Brand1(L)}$</th>
<th>Painkillers</th>
<th>Coffee</th>
<th>Pension plan</th>
<th>Jeans</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{Ad3}$</td>
<td>0.39</td>
<td>0.66</td>
<td>0.54</td>
<td>0.39</td>
</tr>
<tr>
<td>$A_{Ad1(G)}$</td>
<td>0.30</td>
<td>0.66</td>
<td>0.47</td>
<td>0.38</td>
</tr>
<tr>
<td>Sample sizes ($n$)</td>
<td>71</td>
<td>47</td>
<td>51</td>
<td>77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictors of delayed $A_{Brand3}$</th>
<th>Painkillers</th>
<th>Coffee</th>
<th>Pension plan</th>
<th>Jeans</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{Ad3}$</td>
<td>0.41</td>
<td>0.58</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td>$A_{Ad1(G)}$</td>
<td>0.34</td>
<td>0.61</td>
<td>0.56</td>
<td>0.45</td>
</tr>
<tr>
<td>Sample sizes ($n$)</td>
<td>71</td>
<td>47</td>
<td>51</td>
<td>77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictors of delayed $A_{Brand1(G)}$</th>
<th>Painkillers</th>
<th>Coffee</th>
<th>Pension plan</th>
<th>Jeans</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{Ad3}$</td>
<td>0.35</td>
<td>0.39</td>
<td>0.52</td>
<td>0.32</td>
</tr>
<tr>
<td>$A_{Ad1(G)}$</td>
<td>0.28</td>
<td>0.45</td>
<td>0.51</td>
<td>0.39</td>
</tr>
<tr>
<td>Sample sizes ($n$)</td>
<td>71</td>
<td>47</td>
<td>51</td>
<td>77</td>
</tr>
</tbody>
</table>
the single-item reliability. Note that the estimated single-item reliability is a minimum value (Wanous et al. 1997).

Applying the Wanous and Reicher's (1996) approach to the B&R data found that the estimated reliabilities for the single-item measures used in the analysis were the following: 0.97 for AAd1(G), 0.95 for ABrand1(G), and 0.93 for ABrand1(G). These reliabilities are very high according to accepted standards and higher than what scales in marketing and psychology normally attain (Peterson 1994; Peterson and Kim 2013). In a next step, the estimated reliabilities, together with reliabilities for the multiple-item measures, were used to correct the correlations in the original B&R analysis for attenuation (Table 2). The average difference, across the four ads, between the corrected multiple-item correlation and the corrected corresponding single-item correlation was 0.04 (ABrand1(L) as dependent variable), 0.02 (ABrand3), and 0.00 (ABrand1(G)). Thus, the differences between the corrected multiple- and single-item correlations were negligible. (Note that more conservative estimates of the single-item reliabilities would have reduced the differences in the adjusted correlations even further. For example, assuming that the reliabilities are 0.89 will yield single-item correlations that are equal to or higher than the multiple-item correlations. Similar increases in the adjusted correlations would be obtained if it were assumed that the correlation between two error free measures of the same construct is lower than 1.0.)

3 Critique of Kamakura’s (2014) study

In his article, Kamakura (2014) presents results from an analysis of the predictive validity of two constructs (concerns about weight, WEIGHT, and concerns about eating natural, NATURAL). These two constructs were used to predict consumers’ share of low-fat milk and the share of organic milk measured with purchase data. The predictive validity was analyzed by correlating multiple-item measures of the constructs, corrected for attenuation, with the purchase share data. The (corrected) correlations for the multiple-item measures were then compared with the correlations for the individual items that made up the multiple-item measures. The comparison showed that the correlation between the multiple-item measures were consistently higher than the correlations for the individual items, and Kamakura (2014) concludes in the abstract of the article that the “multiple-item scales consistently out-perform their single-item equivalents.”

Unfortunately, there is no definition of the WEIGHT and NATURAL constructs in Kamakura (2014), so it is not possible to know exactly what the items in the two scales attempt to measure. The following discussion is based on “reverse-engineering” by attempting to infer the construct definitions from the items used to measure them. This assumes that the items were chosen because they were considered to be valid indicators of the construct. (An alternative approach would be to assume construct definitions on the basis of the construct labels and descriptions, and to evaluate the measure in relation to the assumed definitions. This approach was not chosen as it is much more open to

---

2 The formula for correction of attenuation is 
\[ \hat{r}_{12} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}} \], where \( \hat{r}_{12} \) = the expected correlation between two perfectly reliable variables, \( r_{12} \) = correlation between variables 1 and 2, and \( r_{11} \) and \( r_{22} \) the reliabilities of variables 1 and 2. If \( \hat{r}_{12} \) is assumed to equal 1.0, then \( r_{22} = \frac{r_{12}^2}{r_{11}} \).
interpretation than the “reverse-engineering” approach.) Four of the five items in the WEIGHT scale relate to behavior (“Pay attention to figure,” “Pay attention to low fat nutrition,” “Worry about weight,” and “Pay attention to calories”) and meet most of the criteria for formative constructs listed by Jarvis et al. (2003, p. 203, note that a construct does not have to meet all criteria to be classified as formative). For example, dropping one item would change the domain of the construct and the items are likely to have different antecedents and consequences. If WEIGHT is formative, which seems highly likely given the items, measurement error on the item level is irrelevant (Diamantopoulos and Winklhofer 2001; Jarvis et al. 2003) and the correction for attenuation is not applicable.

Moreover, judging by the items that make up the measure of the construct, WEIGHT is not a doubly concrete construct as defined by B&R. A construct is doubly concrete if “…the construct is such that in the minds of raters (e.g., respondents in a survey), (1) the object of the construct is ‘concrete singular,’ meaning that it consists of one object that is easily and uniformly imagined, and (2) the attribute of the construct is ‘concrete,’ again meaning that it is easily and uniformly imagined” (B&R, p. 176). (As mentioned in the Introduction, typical examples of doubly concrete constructs are A_Ad, A_Brand, and Pi_Brand. To these could be added other marketing examples such as celebrity attractiveness, celebrity expertise, and brand preference.) While the object in the WEIGHT construct, perhaps, is concrete, the attribute is not. The items describe several widely different behaviors and clearly do not represent an attribute that “…has 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…” (Rossiter 2002, p. 313). Since the attribute of WEIGHT is not concrete, it falls outside of the boundary conditions for when single-item measures should be used, which are clearly stated in B&R (p. 183): “The single-item recommendation for A_Ad and A_Brand cannot be generalized beyond doubly concrete constructs.”
The second construct in the analysis, NATURAL, is probably not a formative construct, and it could be appropriate to correct this construct for attenuation. However, it is, again judging from the items, not a doubly concrete construct. The four items cover a wide range of preferences and dislikes (“Prefer natural products,” “Prefer to buy additive-free food,” “Dislike preservatives in products,” and “Will not purchase environmentally-unfriendly food”) which clearly suggests that the attribute is not concrete. Moreover, the object referred to in the items alternates between “products” and “food,” which suggests that the object is not concrete either.

Thus, the analysis of the WEIGHT and NATURAL constructs in Kamakura (2014) is not a relevant assessment of the predictive validity of single-item measures: Neither construct is doubly concrete and should not, according to B&R, be measured with a single-item measure. Moreover, it should be noted in this context that B&R do not argue that any single-item measure is better than a multiple-item measure. Single-item measures should be carefully selected and only good single-item measures can be expected to outperform their multiple-item equivalents (see also Bergkvist and Rossiter 2009). Comparing a multiple-item measure to an arbitrary selection of single-item measures is pointless as the comparison only is relevant for items selected a priori, on the basis of content validity, as good items.

4 Costs associated with multiple-item measures

Proponents of multiple-item measures frequently argue as if adding items to a measure comes without costs or trade-offs. Naturally, this is not the case: Using multiple-item measures when single-item measures would suffice is not a free lunch. Adding items to a measure not only increases data collection costs but also entails the risk of harming the quality of the research in different ways.

Using multiple-item measures quickly increases the total of number of items in the questionnaire. Using three items instead of one to measure AAd, ABrand, and other doubly concrete constructs may seem harmless, but the sum total of items of all multiple-item measures may be substantial. Consider, for example, a study of the effects of celebrity endorsements on consumer brand evaluations and purchase intention. Such a study typically includes measures of celebrity attractiveness, celebrity expertise, celebrity liking, celebrity-brand fit, AAd, ABrand, and PIBrand, which are all doubly concrete. If single-item measures are used, the total number of items in the study will be seven. If, on the other hand, the study uses the three-item measures that marketing researchers typically use to measure these constructs the number of items are going to be 21. Naturally, the risk of respondent boredom and fatigue is much greater with 21 items rather than seven. For example, a study by Adigüzel and Wedel (2008) found that a 25 % percent decrease in questionnaire response time, from 8 to 6 min, led to a significant reduction in respondent boredom and fatigue. In addition, there is research showing higher response rates for mail surveys if the questionnaire is shortened (Dillman et al. 1993).

Another concern with using multiple items is self-generated validity, that is, when “…the act of measurement changes the phenomenon under study, producing the thought processes predicted by the theory being tested and quite possibly influencing behavior” (Feldman and Lynch 1988, p. 422). According to Feldman and Lynch
(1988), the risk of self-generated validity is the greatest when multiple-item measures are used, even if the items are interspersed throughout the questionnaire. They also note that self-generated validity is likely to increase internal consistency among items and that high internal consistency should not be attributed only to limited measurement error (Feldman and Lynch 1988, p. 427).

A challenge when developing multiple-item measures for doubly concrete constructs is to come up with synonyms for the attribute that is measured. Naturally, as more items are generated, they are going to be less synonymous with the adjective matching the original attribute of interest (Rossiter 2010). For example, for multiple-item measures of $A_{Ad}$ the researcher needs to generate synonyms of the adjective “like.” The difficulty of doing this is illustrated by a review in Bruner (1998) which shows that multiple-item measures of $A_{Ad}$ over the years have included items such as “fresh/stale,” “fair/unfair,” “valuable/not valuable,” and “sensitive/insensitive,” neither of which can be said to be good indicators of whether a consumer likes or dislikes an ad. If consumers rate their $A_{Ad}$ on a multiple-item measure with items like these, it is most likely that ratings will end up around the mid-point of the scale as items perceived as irrelevant tend to be rated neutral (see Rossiter 2010 for an illustrative example). This points to the risk of measures of doubly concrete constructs becoming less valid as items are added. In fact, a meta-analysis of coefficient alpha in Peterson (1994) found an inverse relationship between the number of items in a scale and the average interitem correlation of the scale. This strongly suggests that adding additional items often has a detrimental effect on the validity of a scale.

In addition, there is the problem of what Osgood et al. (1957, p. 187) calls “concept-scale interaction,” which they define as when “the meanings of scales and their relations to other scales vary considerably with the concept being judged.” In the context of measurement, this applies to the situation when individual items either change their meaning depending on the object that is being evaluated or the relevance of the item varies depending on the object. For example, the item “informative,” which is often used in multiple-item measures of $A_{Ad}$, has been shown to be less relevant to ads for emotional (also referred to as hedonic or transformational) products than in ads for rational (utilitarian or informational) products (Bergkvist and Rossiter 2009). This means that as items are added to a measure, there is a risk that it ends up being less valid for some objects than for others.

5 Conclusion and discussion

A re-analysis of the data in B&R shows that the three main criticisms in Kamakura (2014) can be refuted. First, a marker variable analysis (Lindell and Whitney 2001; Malhotra et al. 2006) demonstrated that the amount of common-method variance in the data is negligible. Second, correlations between the predictors from the original study and criterion variables collected in a separate, delayed study yielded the same results as the original study, showing that the timing of the measurement of the criterion variable in B&R was not an important factor for the results. Third, when the single-item measures in B&R are corrected for attenuation, the correlations with the criterion variables were virtually the same as for the multiple-item measures corrected for attenuation. Thus, the issues raised by Kamakura (2014) do not warrant any change.
in the conclusions from B&R and the recommendation to use single-item measures for doubly concrete constructs, and for reviewers and editors to accept manuscripts appropriately using single-item measures, remains unchanged.

A centerpiece of Kamakura’s (2014) criticism of B&R was that the study did not take measurement error into account and corrected the multiple-item measures for attenuation. However, in practice marketing researchers almost never correct for attenuation when reporting results from their research (with the exception of studies using structural equation modeling which automatically takes measurement error into account). If marketing researchers do not correct for attenuation, this begs the question of what is the relevance of correcting for attenuation when comparing single-item and multiple-item measures? Moreover, and more important, as noted by Nunnally and Bernstein (1994, p. 257), “perfect reliability is only a handy fiction, and results from applying the…formula for correction for attenuation are always hypothetical.” One of the problems associated with using the formula for attenuation to correct correlations, mentioned by Nunnally and Bernstein (1994, p. 257), is that it sometimes provides a poor estimate of the “correct” correlation, which is illustrated by the fact that corrected correlations can take on values greater than 1.00. In fact, Nunnally and Bernstein (1994) caution against indiscriminate use of the formula for attenuation, particularly in the context of predictive validity, and suggest it mainly be used to simulate changes in correlations if reliability is increased.

The analysis of predictive validity using consumer panel attitude and purchase data reported in Kamakura (2014) was not a relevant evaluation of single-item measures. Neither of the two constructs used in the analysis were doubly concrete and, hence, B&R’s recommendation to use single-item measures do not apply to them. In addition, one of the constructs is formative and correction of attenuation cannot be validly applied to its measure. This analysis illustrates the importance of theory and domain definition in measurement. The first step in scale development is definition of the construct (e.g., Churchill 1979; DeVellis 2003) and the definition should be retained when the construct is applied in research. Erroneously treating formative constructs as reflective is common in marketing research (Jarvis et al. 2003), something that would be avoided if construct definitions identified whether the construct is reflective or formative. Moreover, Rossiter (2002, 2011) argues that construct definitions should make clear whether objects and attributes are concrete or abstract, a principle, which if it is adhered to, would reduce the risk of using single-item measures for other than doubly concrete constructs.

Compared to other social science disciplines, marketing appears to be prone to measurement conservatism. The article by B&R has been frequently cited. However, out of more than 250 SSCI citations3 (as of March 2014) only about a third appears in marketing journals. The majority of citations are in other fields such as psychology, management, and health research. The oft-cited article by Wanous et al. (1997) is almost never cited in marketing: Less than 20 (<3 %) of almost 700 SSCI citations (March 2014) can be found in marketing journals. Compared to the number of citations in psychology (>200) and management (>100) journals, the number of citations in

---

3 The number of Google Scholar citations is considerably higher (750+). However, Google Scholar, unlike Web of Science, does not support analysis of citations on the publication level. It seems likely that the share would be similar if the Google Scholar citations were analyzed.
marketing journals is very low. The same difference appears to exist also when it comes to accepting studies using single-item measures in top journals. It is easy to find articles in top psychology journals that have used single-item measures for doubly concrete constructs such as thought confidence (e.g., Tormala et al. 2007), liking (e.g., Norton et al. 2007), preference (e.g., Griskevicius et al. 2010), and purchase intention (e.g., Sundie et al. 2011), whereas similar instances in top marketing journals are more difficult to find. Reviewer bias against single-item measures has been identified as an obstacle to marketing journals accepting groundbreaking and interesting research (Lehmann et al. 2011).

For the future development of the field of marketing, it would be beneficial if there was an open-minded and constructive debate about measurement. This debate needs to go beyond the current focus on measurement error and include a wider range of measurement-related topics. This could be a challenge to some since, as noted by Nunnally and Bernstein (1994, p. 214), “[p]erhaps more has been written about measurement error than even more important topics like validity because the theory of measurement error lends itself so well to mathematical treatment.” Hopefully, a majority of marketing academics can go beyond knee-jerk reactions against using single-item measures and focus on more important topics.

Acknowledgments The author is grateful to Tobias Langner and Saeed Samiee for valuable comments on an earlier version of this manuscript.

References


