Stockholm University

This is a published version of a paper published in *Transfer Werbeforschung & Praxis*.

Citation for the published paper:
"The Importance of Choosing One Good Item for Single-Item Measures of Attitude towards the Ad and Attitude towards the Brand and Its Generalization to All Measures"
*Transfer Werbeforschung & Praxis*, 55(2): 8-18

Access to the published version may require subscription.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:diva-37175

http://su.diva-portal.org
The Importance of Choosing One Good Item for Single-Item Measures of Attitude towards the Ad and Attitude towards the Brand and Its Generalization to All Measures

1. Introduction

Bergkvist and Rossiter’s (2007) article demonstrating the sufficiency of single items for measuring “doubly concrete” constructs in marketing – constructs in which there is a single unambiguous object to be rated on a single unambiguous attribute – left open the question of how to select the best single item from the multiple items commonly used to measure these constructs. Rossiter (2002), in his C-OAR-SE procedure for scale development, emphasized the importance of choosing “one good item” for single-item measures and argued that the best item could be identified a priori by expert judgment of item content. Use of the best single item is important because not only is the best single item the most content-valid measure of the construct but also it should be the most valid predictor of scores on a theoretically relevant outcome, or criterion, construct’s measure. Many researchers, however, would prefer empirical proof of the claim that content validity ensures predictive validity and the purpose of the present study is to provide it.

The present study examines the predictive validity of single items taken from typical multiple-item sets of items used to measure the widely employed constructs of Attitude towards the Ad (A\text{Ad}) and Attitude towards the Brand (A\text{Brand}). Beforehand, the authors nominate the single item for each that they judge to be most content valid. In the study, all single items are subjected to a test of predictive validity by using A\text{Ad} single items to predict A\text{Brand}, and then A\text{Brand} single items to predict brand purchase intention, P\text{Brand}. To increase the generalizability of the findings about the best single item for measuring each construct, the coefficients of predictive validity are examined across four new ads for four different products.
The typical academic researcher’s belief that multiple items are always needed has led to the inclusion of items in multiple-item measures of AAd and ABrand that in fact measure antecedents of the attitude rather than the attitude itself. Examples are inclusion of the items “informative” and “credible” in an AAd scale and inclusion of the items “likable” and “useful” in an ABrand scale. These antecedent attributes are “off-attribute” as far as the focal attribute is concerned and are therefore less content-valid measures of the construct.

When ratings of these slightly “off” attributes are averaged in with ratings of the focal attribute, as they are with multiple-item measures, the resulting score will be “off” the true score. Note that the error here is systematic, due to rating the non-valid items, not random error, so it will not cancel to zero in the sum or average score as is popularly believed. Scores on an “off-attribute” item should reduce predictive validity when the item is used in a multiple-item measure or when used alone.

2. Hypotheses

The hypotheses are stated as a priori nominations of the single best item for each construct, given the conceptual definition of the construct. The rationale for each hypothesis follows it, as well as the theoretical rationales for why other commonly used items investigated in the present study should not be “good” items.

H1: For AAd, the single best item is “like-dislike.”

The most widely accepted conceptual definition of attitude toward the ad, AAd, is “a predisposition to respond in a favorable or unfavorable manner to a particular advertising stimulus during a particular exposure occasion” (see Lutz 1985 and MacKenzie and Lutz 1989) and the main attribute underlying attitude toward the ad is affect, or likability (see Brown and Stayman 1992). Both academic theory (Brown and Stayman 1992) and practitioner theory (Haley and Baldwin 1991) suggest that ad likability is the most important advertising stimulus determinant of attitude toward the brand. If so, the academic construct of attitude toward the ad becomes conceptually equivalent to the practitioner’s predictively well-validated construct of ad liking (for the validation studies, see Haley and Baldwin 1991 and Walker and Dubitsky 1994).

However, AAd is often measured in academic studies with multiple items that represent MacKenzie and Lutz’ (1989) antecedent construct – actually separate specific constructs – of perceptions of the ad. Examples are perceptions that the ad is informative (“informative-uninformative,” e.g., Di-mofte, Forehand, and Deshpandé 2003; Forehand and Deshpandé 2001; Pope, Voges, and Brown 2004) or is effective (“good-bad,” e.g., Coulter and Punj 2004; Darley and Smith 1993; Maclnnis and Stayman 1993). These items are antecedents and are therefore “off-attribute” in terms of content validity as measures of AAd.

The item “pleasant-unpleasant” is also occasionally included in item sets to measure AAd (e.g., Darley and Smith 1995; MacKenzie and Lutz 1989; Zhang and Gelb 1996) but judging an ad as pleasant doesn’t exactly mean that you like the ad. The item “pleasant-unpleasant” is “off-attribute” in terms of content validity and also was shown to be liable to concept-scale (in C-OAR-SE terminology, object-attribute) interaction in the original semantic differential research of Osgood, Suci and Tannenbaum (1957).

H2: For ABrand, the single best item is “good-bad.”

The conceptual definition of attitude toward the brand, ABrand, focuses on brand evaluation (Engel, Blackwell, and Miniard 1995 and see also Fishbein and Ajzen 1975). Indeed some researchers (e.g., Langner, Esch and Rossiter 2005; Lee and Labroo 2004) have adopted the term “brand evaluation” rather than “brand attitude.” The item “good-bad” is the marker item (highest loading) for the evaluation factor in the evaluation-potency-activity semantic differential (Osgood et al. 1957). The purpose of advertising is surely to make the brand be evaluated as “good.”

Perhaps in order to be consistent with their definition of attitude toward the ad, AAd, and ABrand measures by other researchers (e.g., Niedrich and Swain 2003; Samu, Krishnan, and Smith 1999; Wheeler, Petty, and Bizer 2005). It could be argued that “favorable-unfavorable” is too vague a description of the focal attribute of evaluation and that it makes sense in conjunction with some objects but not others (object-attribute interaction). For example, some consumers perceive the Nike brand unfavorably because of the company’s past alleged human rights violations, but this may not produce a poor evaluation of Nike-branded products.

Another item typically included in multiple-item measures of ABrand is “like-dislike” (e.g., Darley and Smith 1993, 1995; Y. Lee and Han 2002; Li, Daugherty, and Biocca 2002). However, a brand could be evaluated as “good” regardless of whether the consumer “likes” it. Also, Williams, Cote, and Buckley (1989) raised the possibility of shared method vari-
ance, which would artificially inflate the correlation between $A_{Ad}$ and $A_{Brand}$ when the same attribute, “liking,” appears in both scales – as it did in MacKenzie and Lutz’ (1989) study and in an earlier study by Stayman and Aaker (1988).

The item “pleasant-unpleasant” is often used as an item to measure $A_{Brand}$ (e.g., Batra and Ray 1986; Grier and Deshpandé 2001; MacKenzie and Lutz 1989). However, in connection with products, as with ads, pleasantness is very likely to produce object-attribute interaction. One could conceive of perfume, for example, as being rated very “pleasant” and evaluated as very “good,” whereas a cough syrup might show a contrary relationship by the well-known phenomenon of consumers not regarding a medicine as “effective” unless it tastes “unpleasant.”

Another such “off-attribute” item is “useful-useless” (e.g., Baker, Honea, and Russell 2004; Batra and Ray 1986; MacInnis and Stayman 1993). “Useful” doesn’t really apply to the evaluation of sensory products. This item is therefore likely to suffer from object-attribute interaction, where the objects in this case are various products.

The conative outcome of $A_{Brand}$, brand purchase intention ($PI_{Brand}$), can be measured with a single item measuring the likelihood attribute: “likely-unlikely.” Occasionally a second item, “certain-uncertain,” is used (e.g., Li et al. 2002; Toorman, Davis, and Lemon 2007; and Urbany, Bearden, Kaicker, and Smith-de Borro 1997) but this item may represent hesitancy to act rather than likelihood of acting and, if so, is “off-attribute” for measuring intention. “Likelihood” (to try) is the appropriate item in the present study (which dealt with new products) but the item “certainty” was also investigated and produced an unhypothesized and cautionary result.

3. Research approach

3.1 Overview

The data characteristics in the present study are similar to those in the Bergkvist and Rossiter (2007) study although new samples of ads and consumers are used, with slightly different items in the measures. The items were selected in accordance with the hypotheses and their accompanying rationales. The item sets include, for each of the constructs $A_{Ad}$ and $A_{Brand}$, one prior-judged content-valid item and three other items used by other researchers in their multiple-item measures that are judged to be less valid.

For the study materials, four ads previously unseen by the participants, advertising four different product brands that were new to them, are used.

The ads and brands are rated on the multiple-item sets for each construct in a within-subjects design, with the order of ads rotated across participants.

3.2 Participants

The participants were undergraduate business students attending classes at an Australian university ($N = 117$). Demographically, there were 57% females and 43% males, ranging in age from 18 to 31 years with an average age of 20.

3.3 Procedure

Each participant was given a questionnaire booklet consisting of a series of four 1-page, color magazine ads with the rating scales pertaining to each ad immediately following it on the next page (the questionnaires were reproduced single-sided). The order of ads was rotated across participants. Before the ratings commenced, participants were told that the questionnaires had ads in different orders, so that they would not be surprised if they noticed other students rating different ads. They were told to look at each ad in the questionnaire booklet as they would when reading a magazine but to rate each in turn without looking back. The procedure took about 15 minutes of class time.

3.4 Materials

The ads were color photocopies of English text ads from foreign magazines. The advertised brands were not available in Australia and were therefore presumed to be new to the participants. The ads were for brands in four different product categories: breakfast cereal and painkillers (“informational” appeals) and beer and wine (“transformational” appeals). These were products that many undergraduate students buy and consume, although individual differences were expected such that a range of brand attitudes and purchase intentions would be observed.

3.5 Measures

The measures in the study are shown in Table 1. Sets of four items were used, with the order of items shown within the sets.

4. Analysis

The analysis was done in reverse order but is reported “forwards” to correspond with the order of the hypotheses. The
reverse order was necessitated by first investigating that the single item “likely-unlikely” should be the dependent variable for PI\textsubscript{Brand} in the A\textsubscript{Brand} → PI\textsubscript{Brand} prediction and then establishing that the “best” single item for A\textsubscript{Brand} should be the dependent variable in the A\textsubscript{Ad} → A\textsubscript{Brand} prediction.

5. Results

For the two items measuring PI\textsubscript{Brand}, an interesting result emerged. The correlations of the “certainty” item with the focal item, “likelihood,” averaged only .34 across the four products. An alternative explanation to the hypothesis that “certainty” is an off-attribute item representing “hesitancy” was proposed by Praxmarer (see acknowledgements). The “uncertain” end-point of the answer scale could be interpreted as the midpoint of the attribute, so that the answer scale would not capture the other half of the attribute, which would end in “certainly would not try” (see Rossiter 2002 regarding the erroneous use of unipolar answer scales for bipolar attributes and vice versa). This would produce a restriction of range for correlations involving the “certainty” item and should caution researchers against using it. Even if this item were used with the correct unipolar end-points of “certainly would not try” (0) and “certainly would try” (6), the very low restricted correlation found here suggests that the item is “off-attribute” for measuring purchase likelihood.

By expert judgment of content validity of the item with the purchase intention question asked of respondents, “likely” was chosen as the best single item for measuring PI\textsubscript{Brand} and was used in the analysis for the second hypothesis. By expert judgment which was confirmed empirically, “good” proved to be the best single item for measuring A\textsubscript{Brand} and was used in the analysis for the first hypothesis.

The analyses for H\textsubscript{1} and H\textsubscript{2} are presented below in the normally hypothesized causal, or “hierarchy of effects,” order.

5.1 Single Item Measures of A\textsubscript{Ad} as Predictors of A\textsubscript{Brand}

Predictive validity is not a matter of seeking the highest correlation between a predictor and the criterion but rather coming closest to the true, or “population,” correlation between the two variables (see Rossiter 2002). Because most variables in the social sciences have multiple causes, an observed bivariate predictive validity correlation that is unusually high should be regarded as suspicious. Unless a good estimate of the true correlation is made, one cannot evaluate the observed correlations. For example, an observed r of .5 would be evaluated as accurate or inaccurate depending on whether the true r is .5 or only, say, .3.

A good estimate of the true correlation between A\textsubscript{Ad} and A\textsubscript{Brand} is obtained from the meta-analysis of such correlations by Brown and Stayman (1992). Adjusted for ads for new products, as used in the present study, the estimated true popula-

![Table 1: MULTIPLE-ITEM MEASURES OF A\textsubscript{Ad}, A\textsubscript{Brand}, AND PI\textsubscript{Brand}](image)
tion correlation between $A_{Ad}$ and $A_{Brand}$ is .54 (from Table 4 in their article, take the correlation of .62 for new product ads and multiply it by .53/.61 to estimate the correlation for a single-item measure of $A_{Ad}$). However, for a within-subjects design as presently used, the Brown and Stayman meta-analysis revealed an apparent “practice” effect from rating multiple ads. The .54 correlation therefore needs to be further adjusted upward (by .65/.53, see their Table 4) in order to compare it with the observed correlations, thus .66, or the observed correlations have to be scaled down (by .53/.65) to be compared with the estimated true population correlation of .54. The former alternative was chosen rather than rescaling the observed correlations. This means that validity coefficients (r) that are significantly lower or higher than .66 are suspect. The observed validity coefficients for the single items measuring $A_{Ad}$ and predicting $A_{Brand}$ are shown in Table 2. These are tested (by Fisher’s r-to-z transformations of the correlations and then a T test, see Ferber 1949) for significance of departure from the estimated true validity coefficient of $r = .66$. The 95% confidence interval for N = 117 (the sample size) is bounded by observed r’s of .54 and .75 and the 99% confidence interval by observed r’s of .50 and .78.

The first pattern to note in the tabled results is that three of the validity coefficients for the a priori nominated best single item for $A_{Ad}$, “like,” do not differ significantly (lying within the conventional 95% confidence interval) from the estimated true validity coefficient of $r = .66$. The other, $r = .53$ for the wine ad, is very close to the 95% lower bound of .54 and does not differ within the broader 99% confidence interval from the estimated true validity coefficient.

The second finding is that the validity coefficients for each of the other items, with one exception in 12 comparisons, are significantly lower than the estimated true r of .66, falling outside the 99%, 2-tailed confidence boundaries. The single item “good,” which is the best single item for measuring $A_{Ad}$ (see Table 3 following), is not the best measure of $A_{Ad}$ because its validity coefficient is always significantly below true despite it possibly benefiting from common-methods inflation. The “antecedent” item, “informative,” is also, as expected, a poor predictor, especially for the (transformational) wine ad. The “off-attribute” item, “pleasant” is also a poor predictor, except for the painkiller ad, where its r of .53 is quite close to the true correlation, possibly because ads for “unpleasant” products may be prejudicially rated as “unpleasant.” The differential predictions across products indicate that both items, “informative” and “pleasant,” suffer scale-concept (object-attribute) interaction.

In summary, only the content-valid item “like-dislike” shows generalizable predictive validity as a single-item measure of $A_{Ad}$.

5.2 Single Item Measures of $A_{Brand}$ as Predictors of $P_{IBrand}$

Again, to properly evaluate the validity coefficients, an estimate is needed of the true population correlation, this time between $A_{Brand}$ and $P_{IBrand}$. A meta-analysis by Ouellette and Wood (1998) provides this estimate. For frequently performed behaviors, the correlation, which is for single-item measures of intention and behavior, is .44 (from their Table 2 and assuming that $A_{Act}$, which they surveyed, is the same as $A_{Obj}$, which it would be for frequently purchased products). In their study no adjustment is available for within- vs. between-subjects designs but transferring the adjustment factor from the Brown and Stayman (1992) study, the adjusted estimate for a within-subjects design as presently used, the Brown and Stayman (1992) study, the adjusted estimate for a within-subjects design is .54. This adjusted estimate was adopted rather than the alternative of scaling down the ob-

<table>
<thead>
<tr>
<th>$A_{Ad}$ Single Item</th>
<th>Cereal ad r</th>
<th>Painkiller ad r</th>
<th>Beer ad r</th>
<th>Wine ad r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like</td>
<td>.67a</td>
<td>.58b</td>
<td>.64a</td>
<td>.53b</td>
</tr>
<tr>
<td>Good</td>
<td>.45c</td>
<td>.48b</td>
<td>.49c</td>
<td>.36c</td>
</tr>
<tr>
<td>Informative</td>
<td>.40c</td>
<td>.38c</td>
<td>.37c</td>
<td>.10*bc</td>
</tr>
<tr>
<td>Pleasant</td>
<td>.41c</td>
<td>.53b</td>
<td>.38c</td>
<td>.31c</td>
</tr>
</tbody>
</table>

Notes: N = 117. All r’s are significantly greater than zero at p < .001 except the r that is asterisked. Within columns, validity coefficients with other than the “a” superscript differ significantly from the estimated true population validity coefficient for the within-subjects design, which is $r = .66$, at $p \leq .05$, 2-tailed (“b”) or $p \leq .01$, 2-tailed (“c”). The a priori selected most content-valid single-item measure of $A_{Ad}$ is “like.”
erved correlations to what they might be if the design had been between-subjects. The observed validity coefficients in the present study for $A_{\text{Brand}}$ predicting $P_{\text{Brand}}$ are shown in Table 3. Again, these are tested for significance of departure from the estimated true validity coefficient of $r = .54$. The 95% confidence interval for $N = 117$ is bounded by $r$’s of .40 and .66 and the 99% confidence interval by $r$’s of .35 and .69.

The observed validity coefficients for the a priori best single item, “good,” for measuring $A_{\text{Brand}}$ are all very close to the estimated true validity coefficient of .54, all lying within the 95%, 2-tailed confidence boundaries.

“Like,” which could be expected to be the second-best single item because it is fairly close in meaning to the $A_{\text{Brand}}$ construct’s focal attribute of “evaluation,” is a valid predictor of $P_{\text{Brand}}$ for three of the products but less so, within the broader 99% confidence interval only, for the other, suggesting some degree of object-attribute interaction. “Like” is a slightly less valid predictor ($r = .40$) for the breakfast cereal product, perhaps because the cereal product was advertised for its nutritional benefits rather than likable taste. On the other hand, it makes sense that the consumer has to “like,” meaning “approve of,” taking painkiller products, drinking beer, or drinking wine before intending to buy these products. Although these results suggest that “like” is the “second best” single item for measuring $A_{\text{Brand}}$, this would make it the same as the best single-item measure of $A_{\text{Ad}}$, thus risking common-methods inflation in a study relating the two constructs, and there is no reason to use it when a better item, “good,” is available.

“Useful,” as expected, is “off-attribute” for evaluation and a poor predictor except, interestingly, for wine, where it predicts well ($r = .60$) and may represent a product category belief among students who intend to buy it that wine is “useful” for “getting high.” “Pleasant” is also off-attribute and shows object-attribute interaction in that it predicted well for only one product, painkillers. It is not too difficult to construct the ex post explanation that painkillers should be “not unpleasant” to take. Nevertheless, the important conclusion is that neither of these two items can be relied on as a single-item measure of $A_{\text{Brand}}$.

In summary, only the content-valid item “good-bad” shows generalizable predictive validity as a single-item measure of $A_{\text{Brand}}$.

### 6. Discussion

Averaging scores on the bad items in with the score on the good item may make the multiple-item measure more “reliable” but only in the meaningless sense, for doubly concrete constructs, of Cronbach’s alpha internal consistency (Rossiter 2002). The inclusion of bad items together with the good item certainly makes the multiple-item measure less valid in terms of content validity. This is because “off-attribute” items add systematic rating error to the rating of the “true” focal attribute — of which there is only one in doubly concrete constructs.

The “off-attribute” criticism can also be leveled at answer alternatives within a single-item measure. Juster’s (1966) purchase intention scale is a prime example, where the 11 answer options refer variously to “certainty,” “probability,” “possibility,” and “chance,” which do not uniformly and unambiguously mean likelihood.

The present study has shown that single items that are less content-valid are also less predictively valid. This constitutes empirical support for the importance of Rossiter’s (2002) theoretical recommendation to use “one good item” to measure basic, doubly concrete constructs. Whether the construct is

---

**Table 3: Validity Coefficients (r) of the A_{\text{Brand}} Single Items as Predictors of P_{\text{Brand}}**

<table>
<thead>
<tr>
<th>A_{\text{Brand}} Single Item</th>
<th>Cereal r</th>
<th>Painkiller r</th>
<th>Beer r</th>
<th>Wine r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>.44(^a)</td>
<td>.56(^a)</td>
<td>.50(^a)</td>
<td>.52(^a)</td>
</tr>
<tr>
<td>Like</td>
<td>.40(^b)</td>
<td>.60(^a)</td>
<td>.49(^a)</td>
<td>.42(^a)</td>
</tr>
<tr>
<td>Useful</td>
<td>.37(^b)</td>
<td>.37(^b)</td>
<td>.38(^b)</td>
<td>.49(^b)</td>
</tr>
<tr>
<td>Pleasant</td>
<td>.30(^c)</td>
<td>.47(^a)</td>
<td>.36(^b)</td>
<td>.35(^a)</td>
</tr>
</tbody>
</table>

Notes: $N = 117$. All $r$’s are significantly greater than zero at $p < .001$. Within columns, validity coefficients with other than the “a” superscript differ significantly from the estimated true population validity coefficient for the within-subjects design, which is $r = .54$ at $p < .05$, 2-tailed (“b”) or $p < .01$, 2-tailed (“c”). The a priori selected most content-valid single-item measure of $A_{\text{Brand}}$ is “good.”
doubly concrete can be determined beforehand, if its concreteness is not obvious, by asking a sample of target audience raters what the object and the attribute mean and checking for high consistency of answers (see Rossiter 2002). The good single item can be chosen beforehand, as here, by expert judgment of item content in relation to the conceptual definition of the construct.

A generalization of the “one good item” recommendation can be made, as it was in the C-OAR-SE procedure (Rossiter 2002). It can be extended to the measurement of lowest-order components of an abstract formed attribute. For example, the “service quality” attribute can be conceptually defined as formed from the well-known SERVQUAL second-order components of “reliability,” “assurance,” “tangibles,” “responsiveness,” and “empathy” (Parasuraman, Zeithaml, and Berry 1994), which are themselves formed from first-order (lowest order) components or “indicators” of each. These first-order components must be doubly concrete and each measured with one good single item (see also Rossiter 2007). It could also be argued, as in C-OAR-SE, that the “one good item” recommendation applies to first-order components of an abstract eliciting attribute, in other words, to “reflective” indicators. For example, the personality trait of “extraversion” has as first-order components various elicited mental or physical activities such as “socializing with others,” and “enjoying loud music” (Eysenck 1967). Again, the first-order components must be doubly concrete and each measured with one good single item.

A construct consists of (1) an object, including its constituents or components, (2) an attribute, including its components, and (3) a rater entity (Rossiter 2002). It is important to realize that items can be “off-object” and not only “off-attribute” as in the present study. In the present study, the objects for the AAd items were the ads that the respondents had just viewed, and the objects for the ABand items were the brand representations in those ads, which were iconic packages or logos rather than mere brand names as used in most studies (see Rossiter and Percy 1997). These objects are concrete singular and the items faithfully refer to them. In the case of an abstract collective object, such as “beer,” concrete constituent objects – actual brand labels of representative types of beer – must be used to validly capture the object of the construct. Similarly, in the case of an abstract formed object, for example a consumer value such as “collectivism,” the most common concrete components of the object must be identified and each represented by a single item part; the other part of the single item is the attribute, which in the example of values, is “importance in my daily life” (Rokeach 1973). Note that the first-order objects in a higher-order construct, such as the physical constituents that make up a particular branded website as an object in the construct of “website quality as rated by visitors” (see Rossiter 2007) must also be represented concretely as single item parts and together with the attribute item part must form one good item.

The “one good item” recommendation therefore applies beyond basic constructs and in fact applies to all constructs when they are defined “down” to their concrete components. The first-order components have to be “doubly concrete” in that each has a single clearly understood object as well as a single clearly understood attribute. A content-valid single-item measure of each concrete component of an abstract, or so-called “multidimensional,” construct should produce a maximally content-valid multiple-item measure. This measure should also prove to be the most predictively valid when the abstract construct is used to predict a theoretically relevant criterion construct – which should also be measured with one or more content-valid single items.

The single item must also incorporate an appropriate answer format in order for it to be content-valid (see Rossiter 2002). The importance of the correct answer format was evidenced in the present study by use of the item “uncertain-certain” to measure PIBrand; it should have been “certainly will not…certainly will” so as to cover the full spectrum of probability rather than just half of it. For measuring beliefs (about attributes, or about attribute components in abstract constructs) and intentions (such as PIBrand) the answer scale must be unipolar. For measuring attitude (such as AAd) or evaluation (such as ABand) the answer scale must be affectively bipolar. In neither case should the Likert (“disagree-agree”) item and answer format be used because its answer scale does not refer to the attribute of the construct but rather to a second attribute (agreement) and it will produce ambiguous scores.

As also discussed in the C-OAR-SE procedure, the number of answer categories or scale points is a further issue bearing on the content validity of the item (see Rossiter 2002). A final issue is the way the item is scored, either as a single item or as part of an index or scale of multiple items (see Rossiter 2002, 2008). Sometimes conjunctive (profile) scoring is most valid (e.g., for scoring the abstract formed attribute of “creativity,” which requires high originality and high usefulness); sometimes multiplicative scoring is most valid (e.g., for the multiattribute index of attitude where belief scores are multiplied by attribute evaluation scores in studying attitude formation); and in the remainder of cases simple addition or averaging suffices.
The one good item recommendation therefore assumes good content validity of the question, the answer format and answer categories, and the scoring method. The content validity of all these elements can be established – before use, not after – by expert judgment, aided in difficult cases by pre-test input from cognitive interviews with a modest-sized sample (n = 20 is usually enough) of rater-entity respondents.

It was emphasized in the present article and in C-OAR-SE (Rossiter 2002) that to establish predictive validity, the researcher must first make an estimate of the true (population) validity coefficient. If this estimate is not made, observed validity coefficients cannot be evaluated. In the bivariate case the validity coefficient is r and in the multivariate predictor case it is R, not R² (see Cohen 1977). If no good estimate is available from a previous meta-analysis, then a new estimate of the true validity coefficient can be derived from a new study with content-valid measures (all ultimately based on good single-item measures). Noting that object, attribute, and rater entity are the three necessary defining elements of constructs according to the C-OAR-SE procedure, the new study must be conducted over a representative sample of objects of the constructs using a representative sample of raters.

Empirical tests of the theoretical generalization of the “one good item” recommendation to multiple-item measures – which are, after all, simply aggregations of single items – will provide many opportunities for future research. Such tests will also contribute importantly to our marketing knowledge.

References


