2007

The predictive validity of multiple-item versus single-item measures of the same constructs

Lars I. Bergkvist
University of Wollongong, lars@uow.edu.au

John Rossiter
University of Wollongong, jrossite@uow.edu.au

The predictive validity of multiple-item versus single-item measures of the same constructs

Abstract

This study compares the predictive validity of single-item and multipleitem measures of attitude toward the ad (AAd) and attitude toward the brand (ABrand), which are two of the most widely measured constructs in marketing. The authors assess the ability of AAd to predict ABrand in copy tests of four print advertisements for diverse new products. There is no difference in the predictive validity of the multiple-item and single-item measures. The authors conclude that for the many constructs in marketing that consist of a concrete singular object and a concrete attribute, such as AAd or ABrand, single-item measures should be used.

Publication Details

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.

LARS BERGKVIST and JOHN R. ROSSITER*

This study compares the predictive validity of single-item and multiple-item measures of attitude toward the ad (A_{Ad}) and attitude toward the brand (A_{Brand}), which are two of the most widely measured constructs in marketing. The authors assess the ability of A_{Ad} to predict A_{Brand} in copy tests of four print advertisements for diverse new products. There is no difference in the predictive validity of the multiple-item and single-item measures. The authors conclude that for the many constructs in marketing that consist of a concrete singular object and a concrete attribute, such as A_{Ad} or A_{Brand}, single-item measures should be used.

The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs

In his extremely influential *Journal of Marketing Research* (*JMR*) article, Churchill (1979, p. 66) states the following: "In sum, marketers are much better served with multi-item than single-item measures of their constructs, and they should take the time to develop them." In making that recommendation, Churchill followed the tradition of psychometrics for the measurement of abilities and traits (e.g., Guilford 1954; and especially Nunnally 1978). In the 28 years since Churchill's article, academics have increasingly used multiple items to measure every marketing construct. To be more technically precise, they have used multiple items to measure the *attribute* of the construct (e.g., attitude, quality, liking), as distinguished from the *object* of the construct (e.g., a company, a brand, an ad). In his C-OAR-SE procedure for scale development, Rossiter (2002) proposes that if the object can be conceptualized as concrete and singular, it does not require multiple items to represent it in the measure, and if the attribute can be conceptualized as concrete, it does not require multiple items either. However, Churchill's article, as well as Peter's (1979) *JMR* article on multiple-item reliability, has influenced the measurement of marketing constructs to such an extent that it is virtually impossible to get a journal article accepted in marketing unless it includes multiple-item measures of the main constructs (again, technically, multiple items representing the attribute of the construct). The use of multiple-item measures is also encouraged by the growing popularity of structural equation modeling (e.g., LISREL), a class of statistical techniques for which multiple-item measures are the norm no matter what type of construct is being measured (see, e.g., Anderson and Gerbing 1988; Baumgartner and Homburg 1996).

Among practitioners, use of multiple-item measurement of the same constructs that academic researchers measure is uncommon. Practitioners seem to favor single-item measures, not on theoretical grounds to which practitioners are unlikely to have been exposed, such as those proposed by Rossiter (2002), but on the practical grounds of minimizing respondent refusal and cost. Prominent among these same constructs are attitude toward the ad, or A_{Ad}, which practitioners call "ad liking," or L_{Ad}, and attitude toward the brand, which practitioners and many academics call "brand attitude," symbolized as A_{Brand} or sometimes A_b. These popular constructs are the focus of the current study.

This study adopts the most important criterion for decision-making purposes, predictive validity (Aaker et al. 2005), and demonstrates that single-item measures of these constructs are equally as valid as multiple-item measures. Equal predictive validity means that theoretical tests and empirical findings would be the same if single-item measures were to be used in place of the usual multiple-item

*Lars Bergkvist is Assistant Professor of Marketing, Division of Business and Humanities, UNSW Asia, Singapore (e-mail: l.bergkvist@unswasia.edu.sg). When this study was conducted, he was Senior Lecturer in Marketing, School of Management and Marketing, University of Wollongong, Australia. John R. Rossiter is Research Professor of Marketing, Marketing Research Innovation Centre, University of Wollongong, Australia (e-mail: john_rossiter@uow.edu.au). He is also Visiting Professor of Marketing, Rotterdam School of Management, Erasmus University, the Netherlands. The authors gratefully acknowledge research support from the Association of Swedish Advertisers, the Swedish Newspaper Publishers' Association, and the Advertising Association of Sweden.

To read and contribute to reader and author dialogue on JMR*, visit http://www.marketingpower.com/jmrblog.*

ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. XLIV (May 2007), 175–184

measures for these constructs. We review arguments for and against multiple-item measures and then offer our hypotheses.

ARGUMENTS FOR MULTIPLE-ITEM MEASURES

On what grounds do academics argue that multiple-item measures are better than single-item measures? One theoretical argument that is popular among academic researchers, and can be sourced to Churchill (1979) and the other widely cited *JMR* article by Peter (1979) on reliability, is that multiple-item measures are inherently more "reliable" because they enable computation of correlations between items, which, if the correlations are positive and produce a high average correlation (i.e., a high coefficient alpha), indicate the "internal consistency" of all the items in representing the presumed underlying attribute. This "reliability" argument needs to be qualified (see Rossiter 2002). First, alpha should never be used without first establishing the unidimensionality of the scale (Cortina 1993); this can be investigated by factor analysis or, more safely, by Revelle's (1979) coefficient beta, which is a good test of unidimensionality. Given unidimensionality, alpha is actually an indicator of the validity of the set of items for measuring a certain type of attribute—specifically, an "eliciting" attribute (see Rossiter 2002), of which the main exemplars are personality traits (and corresponding short-term personality states) and abilities. Alpha reliability (or alpha validity) is not relevant for the other two types of attributes—"concrete" attributes (the type of attribute in the constructs A_{Ad} and A_{Brand} in the current study) and "formed" attributes, such as social class (a composite attribute that sums demographic prestige ratings). If the attribute of the construct is concrete, alpha reliability is not a relevant criterion for evaluating the measure, because multiple items to measure the attribute are not necessary. A logical argument against the necessity of high alpha reliability is made by Gorsuch and McFarland (1972), who point out that an unreliable measure cannot form a relationship that yields high predictive validity, and therefore, a single-item measure that is equally predictively valid as a multiple-item measure must be regarded as sufficiently reliable to replace that measure. Cronbach (1961, p. 128) also states, "If predictive validity is satisfactory, low reliability does not discourage us from using the test," meaning here the predictor measure. On the basis of these arguments, it is concluded that reliability need not be considered if the single-item measure demonstrates predictive validity equal to that of the multiple-item measure.

A second theoretical argument for multiple-item measures is that a multiple-item measure captures more information than can be provided by a single-item measure. This argument comes in two forms. One argument for a multiple-item measure capturing more information than a single-item measure is that a multiple-item measure "is more likely to tap all facets of the construct of interest" (Baumgartner and Homburg 1996, p. 143). However, the presence of facets, or components, in an attribute or in an object means that the construct cannot be classified as a concrete attribute of a concrete singular object, in Rossiter's (2002) terminology. Thus, this argument is not relevant for the current study, because our focus is on doubly concrete constructs.

The other form of the "more-information" argument stems from the notion that the multiple-item measure offers more response categories than the single-item measure. It is important to emphasize that it is not the multiple items that are important but rather the number of categories, or "length," of the response scale. Multiple items de facto provide a (potentially) more discriminating response scale than one item. For example, an A_{Ad} measure with three items (i) with seven-point response scales (r) has 343 (r^i) unique response patterns and 19 possible total scores ($i \times r - [i - 1]$). The relatively large number of total scores makes it possible to "make relatively fine distinctions among people" (Churchill 1979, p. 66) or, along the same lines, to categorize people into a large number of groups (Nunnally and Bernstein 1994, p. 67). This is a valid argument as long as the typical respondent can discriminate a large number of categories of the attribute (Viswanathan, Sudman, and Johnson 2004). It follows that a multiple-item predictor measure should show an increased correlation with the criterion measure; that is, it should exhibit higher predictive validity. Moreover, by the same argument, it follows that the correlation of a single-item predictor with a multiple-item criterion should be higher than if both are single-item and that the correlation of a multiple-item predictor with a multiple-item criterion should be highest of all. This form of the more-information argument leads to three hypotheses.

H_1: The correlation between a multiple-item predictor and a single-item criterion is higher than the correlation between a single-item predictor and the same criterion.

H_2: The correlation between a single-item predictor and a multiple-item criterion is higher than the correlation between the same single-item predictor and a single-item criterion.

H_3: The correlation between a multiple-item predictor and a multiple-item criterion is higher than the correlation of any two measures that involves a single-item measure.

ARGUMENTS FOR SINGLE-ITEM MEASURES

Practitioners' preference for single-item measures is not theoretically based but rather is practical, in that single-item measures minimize respondent refusal and reduce data collection and data-processing costs. The only theoretical (versus empirical) argument for using a single-item measure rather than a multiple-item measure has been proposed by Rossiter (2002) in his C-OAR-SE procedure for scale development. Rossiter argues that a single-item measure is sufficient 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. In both cases, "easily and uniformly imagined" is a criterion taken from Wittgenstein's (1961) "picture theory" of language. According to expert judgment based on the C-OAR-SE procedure, A_{Ad} (or L_{Ad}) and A_{Brand} are two such constructs.

An empirically based argument for the use of a single item can be made for measures in which the multiple items representing the attribute (in the answer part of the item) are synonyms, or intended synonyms (more precisely, synonymous adjectives). An extreme example is Zaichkowsky's (1985) well-known measure of personal involvement (as a construct, this refers to personal involvement with some object, such as a product category or an advertisement). Her measure uses 20 bipolar pairs of synonymous adjectives to

measure the attribute of "involvement." Two further examples are represented by the "attitude" attribute in the A_{Ad} and A_{Brand} constructs, as measured by academics. On the basis of a prior study by Stuart, Shimp, and Engle (1987), Allen (2004) uses eight pairs of synonymous adjective items to measure A_{Ad}. This is an exceptionally large number of items; it is more typical for academics to use three or four synonymous items to measure A_{Ad} or A_{Brand} because these are enough to produce a high coefficient alpha. The empirical argument for using a single item for such measures arises because Drolet and Morrison (2001) find that increasing the number of synonymous-answer items produces a frequent problem; specifically, the larger the number of synonymous items the researcher attempts to generate, the greater is the chance of including items that are not proper synonyms of the original attribute descriptor. Moreover, the nonsynonyms are unlikely to be detected. Drolet and Morrison find that respondents were more likely to respond in the same way to an unequivalent (nonsynonymous and, therefore, not content-valid) item as to the other items in a scale when the number of items was increased. Drolet and Morrison include "unfamiliar/familiar" as an unequivalent item in a battery of A_{Ad} items and find that the mean absolute difference in ratings on the equivalent and unequivalent items decreased as the number of items was increased from two items to five items to ten items (one of which was the unequivalent item). The mean difference between the first item and the unequivalent item decreased by approximately 20% when going from two to five items and by approximately 38% when going from two to ten items. Their results suggest (at least for the type of item that is a poor attempt to add an answer synonym) that the addition of more good items hides the presence of bad items. If the bad items are positively correlated with the good items, coefficient alpha increases, an outcome that usually forestalls the researcher from searching for bad items. Paradoxically, the bad items could increase the predictive validity of the multiple-item measure if variation in scores on the new item is correlated with the variation in scores on the criterion, which is likely if the bad item is also another predictor of the criterion. Moreover, Drolet and Morrison apply theory from the field of experts' forecasts to estimate mathematically the informational value of additional items in a scale (see also Morrison and Schmittlein 1991). Using the assumption of moderately correlated errors, they show that additional items provide little information; two items with an error correlation of .60 provide the equivalent of 1.25 independent items, and an infinite number of .60 correlated items provide as much information as only 1.67 independent items. They conclude that one or two good items can outperform a scale with multiple items if the multiple items have moderate or high error correlations, which they are likely to have if they are presented together. Drolet and Morrison's argument is entirely mathematical, and they do not test the additional informational value of questionnaire items empirically. (In the current study, we empirically address the value of additional information by investigating whether multiple items increase predictive validity. If multiple items add information, a multiple-item predictor measure should predict the criterion scores with smaller deviations, resulting in a higher r and R-square.) Because of the problems of systematic errors in the scores obtained from multiple-item measures and because of their mathematical demonstration, which shows that additional items beyond the first do not add much to the prediction of outcomes, Drolet and Morrison recommend the use of single-item measures. However, it must be cautioned that their recommendation would apply only to constructs that constitute the most basic classifications of object and attribute—namely, doubly concrete object and attribute (Rossiter 2002).

Another empirical argument for single-item measures can be derived from the desire to avoid common methods bias. Common methods bias occurs when the correlation between two or more constructs is inflated because they were measured in the same way (see, e.g., Williams, Cote, and Buckley 1989). Common methods bias could occur within the multiple items of a multiple-item measure and, incidentally, would inflate its coefficient alpha. For example, the correlation between A_{Ad} and A_{Brand} is likely to be inflated if each construct is measured with several identical-format items (e.g., "semantic differential" items) rather than a single identical-format item. Common methods bias could also inflate the correlation between two single-item measures if an identical format is used for both. Finally, again between two single-item measures, common methods bias could inflate their correlation if the same (versus different) attribute descriptors are employed (e.g., "good/bad" for A_{Ad} and "like/dislike" for A_{Brand}). Thus, we derive three hypotheses about common methods bias.

H_4: The correlation between two constructs is higher if these constructs are measured with multiple identical-format items than if each construct is measured with a single identical-format item.

H_5: The correlation between two constructs is higher if these constructs are measured with single-item measures with identical formats than if they are measured with single-item measures with different formats.

H_6: The correlation between two constructs is higher if these constructs are measured with single items that employ the same attribute descriptors than if they are measured with single items that employ different attribute descriptors.

Table 1 summarizes the arguments against and for multiple-item measures and how to test these arguments, if it is possible to do so. There are two important empirical tests that can be conducted (see Argument 3 in both lists in Table 1). One test is based on the "discriminability" argument for multiple items and is a test of predictive validity (H_1, H_2, and H_3). The other test is for alternative potential sources of common methods bias (H_4, H_5, and H_6).

Assessing Validity

How can a researcher decide whether a single-item measure of a given construct is as valid as a multiple-item measure of the same construct? Rossiter's (2002) C-OAR-SE procedure states that it is completely a matter of the content validity of the alternative measures. Although consumer open-ended interviews may be needed as input, content validity is ultimately decided by expert judges, and no quantitative research or statistical test apart from interjudge agreement can decide the validity question. However, the expert judgment method is not an option in the current study, because it examines existing measures, for which content validity judgments made *ex post* offer no more than face validity, which is not a valid type of validity, because it

Table 1
ARGUMENTS FOR AND AGAINST MULTIPLE-ITEM MEASURES, AND HOW TO TEST THEM

Arguments for Multiple Items	*Comment*	*How to Test Them*
1. Increases reliability by allowing for calculation of coefficient alpha.	Applies to all constructs according to Churchill's (1979) paradigm. Applies to eliciting attributes according to Rossiter's (2002) paradigm but not to concrete or formed attributes.	Cannot test whether attributes are concrete or formed. Must be decided by expert judgment. For eliciting attributes, coefficient alpha can be calculated when unidimensionality has been established.
2. Necessary if object is abstract or attribute is abstract.	Both Churchill's (1979) and Rossiter's (2002) paradigms accept this, though the terms here are Rossiter's. However, Churchill's paradigm argues that multiple items are necessary for *all* constructs to "tap all facets of a construct." This is not accepted by Rossiter's paradigm (see Argument 1 in "Arguments Against Multiple Items").	Cannot be tested. Decided by expert judgment.
3. Capable of recording greater discrimination (when this is desirable) in categories of the attribute by increasing the number of categories in the answer scale.	Both accept this, though Rossiter (2002) would argue that a single item could be made equally discriminating by increasing the number of categories in the answer scale.	Compare the prediction when the predictor and criterion variables are measured with a multiple-item scale (e.g., three seven-point items, providing 19 possible answer categories) versus a single item (e.g., one seven-point item, 7 categories). If the greater discrimination or "more-information" argument is correct, predictive validity should be highest for multiple-item measures of predictor and criterion, lower for either measured with multiple items, and lowest when predictor and criterion are single-item measures.
Arguments Against Multiple Items	*Comment*	*How to Test Them*
1. Multiple items are unnecessary (not valid) if object is concrete singular and attribute is concrete.	The current study uses A_{Ad}, $Beliefs_{Brand}$, and A_{Brand}. In Rossiter's (2002) framework, each has a concrete singular object (the advertisement or the branded product), and the attributes (belief or attitude) are concrete, so a single item should suffice for each.	Cannot be tested. Decided by expert judgment.
2. Additional items run the risk of tapping into another predictive attribute.	According to Rossiter (2002), this is likely if the items are attempted synonyms of the original attribute.	Break out items of multiple-item scale as independent predictors (stepwise): Additional items should not increase R-square (adjusted) significantly if they tap the same attribute. (Note that a "no-difference" result on the test of Argument 3 in the "Arguments for Multiple Items" section would also be evidence that no other attribute has been tapped.)
3. Common methods bias in predictor and criterion.	Common methods bias could inflate the correlation of a single-item predictor with a single-item criterion but less so than for a multiple-item predictor, a multiple-item criterion, or especially if both are multiple-item measures. With a single-item predictor and single-item criterion, common methods bias could occur with use of an identical format or identical descriptor adjectives for the attribute.	Compare the prediction when the predictor and criterion variables are measured with multiple items of the same type of measure (e.g., semantic differential scale items) versus when the predictor and criterion variables are measured with single items of the same type of measure; compare the prediction when the predictor variable uses the same type of measure as the single-item criterion variable versus when the types of measure differ (e.g., labeled bipolar response scale for the predictor and semantic differential scale item for the criterion variable); compare the prediction for single-item identical adjectives in the predictor and criterion versus single-item parallel adjectives.

does not reveal items that were considered and rejected and therefore does not show how the face-valid items were selected (Rossiter 2002).

The usual psychometric method for comparing validities is to examine how well each measure predicts some relevant outcome measure (called "concurrent validity" when both measures are taken in the same study and "predictive validity" when the criterion measure is delayed, but the term "predictive validity" is commonly used to refer to both situations). On the one hand, Rossiter (2002) raises an objection to predictive validity (see also Borsboom, Mellenbergh, and Van Heerden 2004) because the purpose should not be to maximize the prediction (maximize the magnitude of r) but to match the true correlation (the magnitude of the

population R) between the predictor and the criterion. The true correlation, R_{xy}, will usually be considerably smaller than 1.0 because most outcomes have multiple causes, and in the social sciences, such correlations would be suspicious if they were larger than .6 (Cronbach 1961). On the other hand, if the two predictors being compared are two (or more) measures of the same construct (and, thus, the same attribute), this objection seems less sustainable because though the true correlation is not known, it can be fairly safely assumed that the higher of the correlations is closer to the truth.

In the current study, we employ bivariate correlation analysis and multivariate regression analysis to compare the abilities of single-item and multiple-item measures of attitude toward the ad to predict single-item and multiple-item measures of attitude toward the brand. First, if the multiple-item "greater-discriminability" argument is correct, multiple-item measures of the independent variable, the dependent variable, or both should result in larger validity coefficients, r, and greater accounted-for variance in regressions, R-square, than single-item measures. Second, if the common-methods-bias argument is correct, multiple-item measures should produce unduly inflated predictions. This should also be the case for single-item measures that use the same type of answer scale or the same adjective for the predictor measure and the criterion measure.

RESEARCH APPROACH

Overview

The data in this article come from consumers' responses to four advertisements for four different products. We pretested the advertisements in traditional ad tests (copy tests). We balanced the order of the advertisements tested by rotation. Each participant completed multiple-item and single-item measures of the main ad-test variables so that the comparison of methods of measurement was based on a within-subjects rather than a between-subjects design.

Participants

The participants were first- and second-year undergraduate business students who agreed to participate in "a study about marketing." Participants were offered a free lunch during the copy test, a Red Cross lottery ticket, and the chance to win cinema tickets or gift vouchers for the student bookshop. Overall, 92 participants completed the ad tests, but the cell sizes differed for the four advertisements because of the screening out of participants for whom the particular product category was not personally relevant.

Procedure

Some weeks before the ad test, we carried out a preliminary survey to measure the personal relevance of the product categories in the study. We measured personal relevance by asking about purchase intentions, purchase, and usage for the four product categories, though we omitted the purchase question for one of the product categories, retirement pension plans, because undergraduate students would not have purchased a pension plan. To avoid sensitizing the participants to the product categories to be used in the study, we asked the same questions about four other product categories in addition to the four categories in the study. We considered a positive answer on at least one of the three relevance questions (i.e., intend to buy, have bought, or have used) the minimum for the product category to be regarded as being relevant to the participant, and the analyses that follow are based on the replies from only those who answered positively to at least one of the relevance questions. The proportion of participants with at least one positive answer ranged from 63% to 95% across the four product categories, as can be observed in the n's in the tables of results.

We carried out the ad test in groups of approximately 25 students (with individual booklets) in a classroom during the students' lunch hour. Each participant had previously been assigned to a group and instructed to attend during his or her lunch break. On arrival, the participants were instructed to sit down, wait, and not look through the booklet in front of them. When the test started, the participants were told that they were going to see four advertisements for brands that were not yet available on the local market, but that the brands would be available in the near future. They were also told that there were no right or wrong answers to the questions that would follow each advertisement and that it was their opinions as consumers, not as students at a business school, that mattered. The importance of answering all the questions in the booklet was also emphasized.

Each advertisement in the booklet was followed by all the questions related to it. We rotated the order of the advertisements to minimize carryover effects (an analysis of variance subsequently demonstrated that the order of advertisements was a nonsignificant variable). The participants could take as much time as they wanted to look at each advertisement. Pretests indicated that three minutes was sufficient time for everyone.

Materials

The advertisements were real advertisements for real brands, but neither the advertisements nor the brands were available in the local market in which the study was conducted. Thus, the advertisements and brands were new to the participants. The advertisements were presented in color on A4-size paper, and the paper and the printing were of magazine quality. The brands in the advertisements came from four different product categories: painkillers, coffee, pension plans, and jeans. We chose the product categories a priori to represent the four quadrants in Rossiter and Percy's (1997) grid; these were low involvement/informational, low involvement/transformational, high involvement/informational, and high involvement/transformational, respectively.

Measures

The ad-test questionnaire contained the same questions for all four advertisements in the study. For each advertisement, the participants rated ad liking (L_{Ad}), attitude toward the ad (A_{Ad}), brand purchase intention (PI_{Brand}), brand attitude (A_{Brand}), and brand benefit beliefs ($Beliefs_{Brand}$), in that order. The ad measures came first, immediately after exposure to the advertisement, and then the brand measures were asked in reverse "hierarchy-of-effects" sequence to prevent attitude and intention being constructed from rated beliefs (Rossiter and Percy 1997). The questionnaire included other measures, such as cognitive responses, which, with PI_{Brand}, were not used in the analysis.

Table 2
MEASURES OF THE MAIN CONSTRUCTS

Construct	*Question*	*Answer Scale*
L_{Ad}	"Thinking about the ad for /BRAND/, which of the following statements best describes your feelings about the ad?"	1. I liked it very much. 2. I liked it. 3. I neither liked it nor disliked it. 4. I disliked it. 5. I disliked it very much.
A_{Ad}	"Below you will find three pairs of adjectives. Indicate how well one or the other adjective in each pair describes how you perceived the ad for /BRAND/."	Good ⊔⊔⊔⊔⊔⊔⊔ Bad Unpleasant ⊔⊔⊔⊔⊔⊔⊔ Pleasant Unfavorable ⊔⊔⊔⊔⊔⊔⊔ Favorable
A_{Brand}	"Below you will find three pairs of adjectives. Indicate how well one or the other adjective in each pair describes your overall feeling of /BRAND/ /PRODUCT CATEGORY/."	Bad ⊔⊔⊔⊔⊔⊔⊔ Good Pleasant ⊔⊔⊔⊔⊔⊔⊔ Unpleasant Dislike ⊔⊔⊔⊔⊔⊔⊔ Like

Notes: We used reversed scoring on the single-item L_{Ad} measure (i.e., 5 = "positive" response). We coded multiple items from 1 to 7 for the "semantic differential" A_{Ad} and A_{Brand} measures (7 = "positive" response). For the single-item measures of A_{Ad} and A_{Brand}, one of the three adjective pairs was selected (see the "Measures" section in the text).

The exact scales used to measure the main constructs in this analysis appear in Table 2. We took single-item measures from the multiple-item measures. Ad liking, L_{Ad1}, where the "1" subscript indicates the number of items, was already a single-item measure and is used by most practitioners (Haley and Baldinger 1991; Walker and Dubitsky 1994). Attitude toward the ad, A_{Ad3}, was a three-item measure used in MacKenzie and Lutz's (1989) study and has been used by many other academic researchers after them. For the single-item measure of attitude toward the ad, A_{Ad1}, we selected the first item, "good/bad," which is labeled $A_{Ad1(G)}$ in the results. Brand attitude, A_{Brand3}, was a three-item measure first used by Gardner (1985) and by many other academic researchers subsequently. To examine common methods bias with single-item measures, we selected the third item, "dislike/like," as the "different" single-item measure of A_{Brand1} and labeled it $A_{Brand1(L)}$; for the "same" single-item measure of A_{Brand1}, we selected the same item as for the single-item measure of A_{Ad1}, "bad/good," which is labeled $A_{Brand1(G)}$.

In addition to the main constructs, beliefs about the most important attributes of each branded product, $Beliefs_{Brand}$, were necessary to measure for the regression analyses. $Beliefs_{Brand}$ consisted of the two to four (depending on the product category) most important attributes as determined by a pretest; we measured belief strengths on unipolar seven-point scales ranging from "a very small extent" (1) to "a very large extent" (7). For each branded product, we combined the belief scores for each attribute into an index. We also examined product-by-product regressions with the beliefs as separate independent variables, and they produced nearly identical R-square values. Thus, we report the results for the indexed beliefs to conserve space.

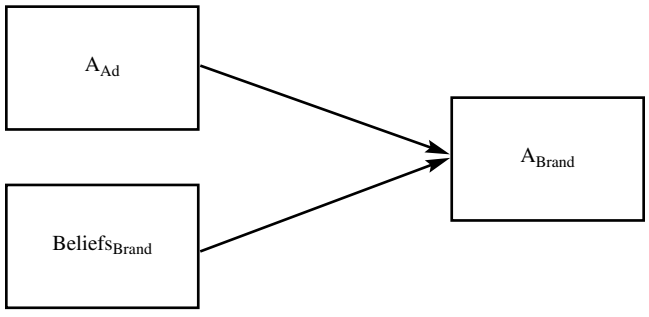
Following classical psychometric procedure (e.g., Cortina 1993), we investigated the multiple-item measures with principal components analysis to check the number of dimensions before we computed their coefficient alpha; we found both A_{Ad3} and A_{Brand3} to be unidimensional. The coefficient alpha for the measures were all good or very good according to accepted psychometric standards; they ranged between .85 and .93 (see, e.g., DeVellis 1991; Nunnally and Bernstein 1994).

ANALYSIS

Overview

The overall aim of the analysis is to compare the predictive validity of multiple-item measures with single-item measures of A_{Ad} and A_{Brand}. Predictive validity is assessed by two methods. One method compares the simple bivariate correlation, r, between the predictor (A_{Ad}) and the criterion (A_{Brand}); r is the usual statistic for reporting a "validity coefficient" in the psychometric test literature for concurrent or, if the criterion is delayed, predictive validity (e.g., Cronbach 1961). The other method is multivariate regression, which compares R-squares. Multivariate regression must also be investigated. Fishbein and Middlestadt (1995), among others, argue that the validity coefficient (correlation) of a predictor with a criterion will be inflated if the model of causes of the criterion is underspecified because the single predictor is likely to include the effects of another causal variable or variables. The most likely causes of A_{Brand} are hypothesized to be A_{Ad} and $Beliefs_{Brand}$ independently (see Figure 1), but if the true causal model includes the indirect causal path $A_{Ad} \rightarrow Beliefs_{Brand} \rightarrow A_{Brand}$, the regression coefficient of A_{Ad} in the reduced model $A_{Ad} \rightarrow A_{Brand}$ will be inflated because it contains and is hiding part of the effect of $Beliefs_{Brand}$ (the mediating part) on A_{Brand}. That is, if the effect of $Beliefs_{Brand}$ is measured and partialled out, the effect of A_{Ad} will be smaller. For the current analyses, the exact underlying theoretical model is without consequence as long as $Beliefs_{Brand}$ is included as a predictor because the statistical solution to dealing with mediating variables is to include both the independent variable and the mediating variable in the regression model (Baron and Kenny 1986).

Figure 1
PRESUMED MODEL OF THE CAUSAL PREDICTORS OF A_{Brand}



We ran the correlations and regressions separately for the four advertisements in the study because aggregated results would be difficult to interpret and might conceal differences between advertisements (or products). We checked all regression models for multicollinearity. None of the models had condition indexes greater than 15 in combination with two or more variance proportions greater than .90 (Hair et al. 1998), which indicates that multicollinearity was not a problem in any of the models. In the analysis, we tested the differences in r and R-square for significance using z tests, following Fisher's normalizing transformation of the correlations (Cohen and Cohen 1975; Howell 1992).

Multiple-Item Versus Single-Item A_{Ad} as Predictors of Single-Item A_{Brand}

The first analysis compares the multiple-item measure of attitude toward the ad, A_{Ad3}. The two single-item measures, the "reduced" measure, $A_{Ad1(G)}$, and the related measure L_{Ad1}, are alternative predictors of the single-item measure of the criterion variable $A_{Brand1(L)}$. Table 3 shows the bivariate validity coefficients, r, and the multivariate validity statistics that represent the accounted-for variance, R-square, with $Beliefs_{Brand}$ in the regression equations.

On the basis of the bivariate validity coefficients, r, we can reject H_1. The single-item measures of attitude toward the ad, $A_{Ad1(G)}$ and L_{Ad1}, were equally good predictors of brand attitude, $A_{Brand1(L)}$, and equally as good as the multiple-item predictor, A_{Ad3} (within advertisements, none of the r's differ significantly, $p > .10$).

The multivariate validity coefficients, R-square, which also appear in Table 3, reveal the same pattern of results, thus rejecting H_1. Confirming the suspicion about omitted causes, the estimated predictive validity of A_{Ad}, as estimated by its standardized regression coefficient, was indeed inflated for three of the four advertised products, with the exception of the pension plan, when we omitted the $Beliefs_{Brand}$ index from the regression model (these analyses are available on request). Therefore, Table 3 reports the R-squares with the index included. The important conclusion is that the causal role of A_{Ad} is unaffected by whether it is measured with multiple items or a single item.

Multiple-Item Versus Single-Item A_{Ad} as Predictors of Multiple-Item A_{Brand}

We repeated the previous analysis using the multiple-item measure of brand attitude, A_{Brand3}, as the criterion variable (Table 4). For the r results, the multiple-item measure of attitude toward the ad, A_{Ad3}, was not a significantly better predictor than the single-item measures $A_{Ad1(G)}$ and L_{Ad1}, and the two single-item measures did not differ significantly from each other (for each advertisement, all *p*s > .10). These results exactly replicated those for the single-item criterion measure, $A_{Brand1(L)}$. The R-square results exactly replicated the results for r, showing that the causal role of A_{Ad} is not affected by whether it is measured with multiple items or a single item. Because the validity coefficient, r, gave the same conclusion as the accounted-for variance statistic, R-square, in all cases, the focus is on r alone for tests of the remaining five hypotheses.

Table 3

VALIDITY COEFFICIENTS (r) AND ACCOUNTED-FOR VARIANCE IN THE MULTIVARIATE REGRESSION[a] (R-SQUARE) FOR MULTIPLE-ITEM AND SINGLE-ITEM MEASURES OF A_{Ad} AS PREDICTORS OF SINGLE-ITEM $A_{Brand1(L)}$

	Advertised Product							
	Painkillers		*Coffee*		*Pension Plan*		*Jeans*	
Predictors of $A_{Brand1(L)}$	*r*	R^2	*r*	R^2	*r*	R^2	*r*	R^2
A_{Ad3}	.75	.58	.77	.72	.68	.48	.68	.58
$A_{Ad1(G)}$	.72	.55	.75	.69	.66	.47	.67	.56
L_{Ad1}	.74	.58	.73	.67	.60	.40	.68	.62
Sample sizes (n)	80		55		59		86	

[a]Regression equations include $Beliefs_{Brand}$.
Notes: All r's are significant at $p < .01$. All regression models are significant at $p < .01$.

Table 4

VALIDITY COEFFICIENTS (r) AND ACCOUNTED-FOR VARIANCE IN THE MULTIVARIATE REGRESSION[a] (R-SQUARE) FOR MULTIPLE-ITEM AND SINGLE-ITEM MEASURES OF A_{Ad} AS PREDICTORS OF MULTIPLE-ITEM A_{Brand3}

	Advertised Product							
	Painkillers		*Coffee*		*Pension Plan*		*Jeans*	
Predictors of A_{Brand3}	*r*	R^2	*r*	R^2	*r*	R^2	*r*	R^2
A_{Ad3}	.80	.69	.75	.80	.72	.52	.65	.61
$A_{Ad1(G)}$	.78	.68	.74	.77	.72	.53	.66	.61
L_{Ad1}	.77	.67	.70	.76	.68	.49	.62	.62
Sample sizes (n)	80		55		59		86	

[a]Regression equations include $Beliefs_{Brand}$.
Notes: All r's are significant at $p < .01$. All regression models are significant at $p < .01$.

Discriminability of Multiple-Item Versus Single-Item Measures

Comparisons of the correlational results in Table 4 with those in Table 3 rule out Churchill's (1979) contention that multiple-item measures are more valid because they can capture greater discrimination of responses because of their provision of more answer categories. This did not hold for either of the constructs in the current study, attitude toward the ad (A_{Ad}) or brand attitude (A_{Brand}). If the hypothesis were true and if consumers really could discriminate finer gradations of attitude than offered by a single seven-point answer scale, the multiple-item predictor measure, A_{Ad3}, should have had the highest correlation with the multiple-item criterion measure, A_{Brand3}. Even the most extreme comparison, $r_{3.3}$ from Table 4 versus $r_{1.1}$ from Table 3, revealed that this was not the case: The correlations were .80 versus .74 for the painkiller advertisement, .75 versus .73 for the coffee advertisement, .72 versus .60 for the pension plan advertisement, and .65 versus .68 for the jeans advertisement. Although the single-item correlations appear to be lower for the two "informational" products, the painkiller and the pension plan, the multiple-item correlations were not significantly higher ($p > .10$). Thus, both H_2 and H_3 were rejected.

Common Methods Bias

Comparisons of the appropriate correlations count against all hypotheses of spuriously inflated correlations due to common methods bias. In the following results, all the relevant comparisons are nonsignificant ($p > .10$). The finding that the multiple-item $r_{3.3}$ correlations (Table 4) between A_{Ad} and A_{Brand} were no higher than the single-item $r_{1.1}$ correlations (Table 3) rejects the notion that repeating semantic differential scales for both measures, at least for two repetitions (i.e., three items), leads to an inflated prediction (H_4). This finding removes a concern with multiple-item measures, at least for up to three items.

For single-item measures, as Table 5 shows, use of the same semantic differential format for the predictor ($A_{Ad1(G)}$) and criterion does not inflate predictions compared with a different-format predictor (L_{Ad1}), measured with a "labeled" answer scale. Thus, the results reject H_5. Finally, again for single-item measures, use of the same adjective descriptor for the predictor and criterion ($A_{Ad1(G)}$ and $A_{Brand1(G)}$ in Table 5) does not lead to inflated prediction compared with different adjectives ($A_{Ad1(G)}$ and $A_{Brand1(L)}$ in Table 3). These results reject H_6.

DISCUSSION

Two of the most widely employed constructs in advertising and consumer research are attitude toward the ad (A_{Ad}) and brand attitude (A_{Brand}). Both constructs are doubly concrete (Rossiter 2002) and therefore should be validly measurable by a single item, even though the overwhelming practice in academic research is to measure them with multiple items. In the current study, for both constructs, the single-item measure demonstrated equally high predictive validity as the multiple-item measure. We obtained this result consistently for all four new product advertisements and for two methods of assessment of predictive validity—the bivariate validity coefficient, r, and the multivariate statistic R-square—when the constructs were included in a causal model. This result fails to support the classic psychometric argument (e.g., Churchill 1979; Nunnally and Bernstein 1994) that multiple-item measures are more valid than single-item measures for all types of constructs. In particular, when multiple-item measures are used to measure doubly concrete constructs, they do not appear to discriminate better by capturing more information, which is the usual justification for their use.

We found no evidence of common methods bias with the multiple-item measures or with single-item measures of the independent variable and dependent variable that use the same measurement format (in this case, semantic differential scales) or the same attribute descriptor (in this case, "good/bad" for both A_{Ad} and A_{Brand}). However, the multiple-item measures in the study consisted of only three items; this is not to say that common methods bias would not inflate predictions when the predictor, the criterion, or both are measured with more items. Although there was no evidence of this in the current study with three items, multiple items may provide rehearsal episodes that might spuriously inflate the prediction (Feldman and Lynch 1988). With more than three multiple items, the number in the current study, this spurious increase in predictive validity may occur.

An important boundary condition on our findings arises from the two constructs in the current study having neither a multicomponent or multiconstituent object nor a multi-

Table 5

VALIDITY COEFFICIENTS (r) AND ACCOUNTED-FOR VARIANCE IN THE MULTIVARIATE REGRESSION[a] (R-SQUARE) FOR MULTIPLE-ITEM AND SINGLE-ITEM MEASURES OF A_{Ad} AS PREDICTORS OF SINGLE-ITEM $A_{Brand1(G)}$

	Advertised Product							
	Painkillers		*Coffee*		*Pension Plan*		*Jeans*	
Predictors of $A_{Brand1(G)}$	*r*	R^2	*r*	R^2	*r*	R^2	*r*	R^2
A_{Ad3}	.70	.62	.70	.76	.63	.40	.53	.46
$A_{Ad1(G)}$	.71	.64	.70	.75	.65	.42	.57	.49
L_{Ad1}	.68	.62	.65	.73	.56	.32	.55	.51
Sample sizes (n)	80		55		59		86	

[a]Regression equations include $Beliefs_{Brand}$.

Notes: All r's are significant at $p < .01$. All regression models are significant at $p < .01$.

component attribute but rather a concrete singular object (the advertisement or the brand) and a concrete attribute (attitude). The single-item recommendation for A_{Ad} and A_{Brand} cannot be generalized beyond doubly concrete constructs. Rossiter's (2002) theory explains why multiple items are needed to measure abstract constructs validly. A construct is "abstract" if (1) the object of the construct comprises two or more components (e.g., the materialism value, which has three components—namely, use of possessions to judge success, centrality of possessions in a person's life, and the belief that possessions lead to happiness; see Richins 2004) or comprises a set of constituent subobjects (e.g., for job satisfaction, aspects of a person's job, such as supervisor, coworkers, job duties, workplace technology, and policies; see Gardner et al. 1998; Locke 1969) or (2) the attribute of the construct is formed from two or more components (e.g., service quality, with its components of reliability, responsiveness, empathy, and so forth; see Parasuraman, Zeithaml, and Berry 1994) or elicits and is reflected in a series of mental or physical activities (e.g., the personality trait of extraversion, which is reflected in risk-taking, gregarious, and energetic activities; see Eysenck 1967). Single-item measures in these two abstract object cases and two abstract attribute cases would be expected to be less valid because the meaning of the object in the single-item questions (e.g., "How important to you is materialism?" "How satisfied are you with your job?") or the attribute in the single-item questions (e.g., "How good is McDonald's service?" "Are you extraverted?") differs far too greatly over raters. Instead, the abstract object must be broken down into concrete components or constituents, each measured with a separate single-item part, and the abstract attribute must also be broken down into concrete components, each measured with a separate single-item part. The object component or constituent item part and the attribute component item part form the item, and multiple items are required. We do not claim that a single-item measure can adequately represent an abstract construct.

The current study can be viewed as an extension of Churchill's (1979) procedure for scale development. Churchill introduced a systematic approach to scale development that has contributed fundamentally to marketing research methodology. He emphasized the importance of theoretical considerations (domain specification) as the first step of his procedure. Our extension to single-item measures is essentially based on, and bounded by, theory. Unfortunately, Churchill's emphasis on theory has received much less attention than his recommendation that marketing researchers should use multiple-item measures. If marketing researchers had been more concerned with the theory of marketing constructs, there would probably have been less mindless use of multiple-item measures in marketing.

Advertisements and brands are probably the two most common objects of study in marketing by both academics and practitioners, and there is no reason that our findings should not generalize to other objects of study in marketing, such as companies, retailers, salespeople, prices, and sales promotions, provided that these objects are concrete and singular. Similarly, attitude is the most widely measured attribute in marketing, and the findings should generalize to other concrete attributes, such as beliefs or perceptions, intentions, and satisfaction. Theoretical tests and empirical findings would be unchanged if good single-item measures were substituted for these constructs in place of commonly used multiple-item measures. Therefore, marketing journals should be willing to accept articles with single-item measures of doubly concrete constructs.

REFERENCES

Aaker, David A., V. Kumar, George S. Day, and Meredith Lawley (2005), *Marketing Research*, The Pacific Rim ed. Milton, Queensland, Australia: John Wiley & Sons.

Allen, Chris T. (2004), "A Theory-Based Approach for Improving Demand Artifact Assessment in Advertising Experiments," *Journal of Advertising*, 33 (Summer), 63–73.

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (May), 411–23.

Baron, Reuben M. and David A. Kenny (1986), "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51 (December), 1173–82.

Baumgartner, Hans and Christian Homburg (1996), "Applications of Structural Equation Modeling in Marketing and Consumer Research: A Review," *International Journal of Research in Marketing*, 13 (April), 139–61.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden (2004), "The Concept of Validity," *Psychological Review*, 111 (October), 1061–1071.

Churchill, Gilbert A. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, 16 (February), 64–73.

Cohen, Jacob and Patricia Cohen (1975), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortina, Jose M. (1993), "What Is Coefficient Alpha? An Examination of Theory and Applications," *Journal of Applied Psychology*, 78 (February), 98–104.

Cronbach, Lee J. (1961), *Essentials of Psychological Testing*, 2d ed. New York: Harper & Row.

DeVellis, Robert F. (1991), *Scale Development*. Newbury Park, CA: Sage Publications.

Drolet, Aimee L. and Donald G. Morrison (2001), "Do We Really Need Multiple-Item Measures in Service Research?" *Journal of Service Research*, 3 (February), 196–204.

Eysenck, Hans J. (1967), *The Biological Basis of Personality*. Springfield, IL: Thomas.

Feldman, Jack M. and John G. Lynch (1988), "Self-Generated Validity and Other Effects of Measurement on Belief, Attitude, Intention, and Behavior," *Journal of Applied Psychology*, 73 (August), 421–35.

Fishbein, Martin and Susan E. Middlestadt (1995), "Noncognitive Effects on Attitude Formation and Change: Fact or Artifact?" *Journal of Consumer Psychology*, 4 (2), 181–202.

Gardner, Donald G., L.L. Cummings, Randall B. Dunham, and Jon L. Pierce (1998), "Single-Item Versus Multiple-Item Measurement Scales: An Empirical Comparison," *Educational and Psychological Measurement*, 58 (December), 898–915.

Gardner, Meryl Paula (1985), "Does Attitude Toward the Ad Affect Brand Attitude Under a Brand Evaluation Set?" *Journal of Marketing Research*, 22 (May), 192–98.

Gorsuch, Richard L. and Sam G. McFarland (1972), "Single Versus Multiple-Item Scales for Measuring Religious Values," *Journal for the Scientific Study of Religion*, 11 (1), 53–64.

Guilford, J.P. (1954), *Psychometric Methods*. New York: McGraw-Hill.

Hair, Joseph F., Rolph E. Anderson, Ronald L. Tatham, and William C. Black (1998), *Multivariate Data Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.

Haley, Russel I. and Allan L. Baldinger (1991), "The ARF Copy Research Validity Project," *Journal of Advertising Research*, 31 (April–May), 11–32.

Howell, David C. (1992), *Statistical Methods for Psychology*, 3d ed. Belmont, CA: Duxbury Press.

Locke, Edwin A. (1969), "What Is Job Satisfaction?" *Organizational Behavior and Human Performance*, 4 (November), 309–336.

MacKenzie, Scott B. and Richard J. Lutz (1989), "An Empirical Examination of the Structural Antecedents of Attitude Toward the Ad in an Advertising Pretesting Context," *Journal of Marketing*, 53 (April), 48–65.

Morrison, Donald G. and David C. Schmittlein (1991), "How Many Forecasters Do You Really Have? Mahalanobis Provides the Intuition for the Surprising Clemen and Winkler Result," *Operations Research*, 39 (May–June), 519–23.

Nunnally, Jum C. (1978), *Psychometric Theory*, 2d ed. New York: McGraw-Hill.

——— and Ira H. Bernstein (1994), *Psychometric Theory*, 3d ed. New York: McGraw-Hill.

Parasuraman, A., Valarie Zeithaml, and Leonard L. Berry (1994), "Alternative Scales for Measuring Service Quality: A Comparative Assessment Based on Psychometric and Diagnostic Criteria," *Journal of Retailing*, 70 (Autumn), 201–230.

Peter, Paul J. (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," *Journal of Marketing Research*, 16 (February), 6–17.

Revelle, W. (1979), "Hierarchical Clustering and the Internal Structure of Tests," *Multivariate Behavioral Research*, 14 (1), 57–74.

Richins, Marsha L. (2004), "The Material Values Scale: Measurement Properties and Development of a Short Form," *Journal of Consumer Research*, 31 (June), 209–219.

Rossiter, John R. (2002), "The C-OAR-SE Procedure for Scale Development in Marketing," *International Journal of Research in Marketing*, 19 (December), 305–335.

——— and Larry Percy (1997), *Advertising Communications & Promotion Management*, 2d ed. New York: McGraw-Hill.

Stuart, Elnora W., Terence A. Shimp, and Randall W. Engle (1987), "Classical Conditioning of Consumer Attitudes: Four Experiments in an Advertising Context," *Journal of Consumer Research*, 14 (December), 334–49.

Viswanathan, Madhubalan, Seymour Sudman, and Michael Johnson (2004), "Maximum Versus Meaningful Discrimination in Scale Response: Implications for Validity of Measurement of Consumer Perceptions About Products," *Journal of Business Research*, 57 (February), 108–125.

Walker, David and Tony M. Dubitsky (1994), "Why Liking Matters," *Journal of Advertising Research*, 34 (May–June), 9–18.

Williams, Larry J., Joseph A. Cote, and M. Ronald Buckley (1989), "Lack of Method Variance in Self-Reported Affect and Perceptions at Work: Reality or Artifact?" *Journal of Applied Psychology*, 74 (June), 462–68.

Wittgenstein, Ludwig (1961), "Entry ca. September 29, 1912," in *Notebooks 1914-1916*, G.E.M. Anscombe and G.H. von Wright, eds. London: Basil Blackwell, 7–8.

Zaichkowsky, Judith Lynne (1985), "Measuring the Involvement Construct," *Journal of Consumer Research*, 12 (December), 341–52.