The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0309-0566.htm

COMMENTARY

Marketing measurement revolution

The C-OAR-SE method and why it must replace psychometrics

John R. Rossiter
Institute for Innovation in Business and Social Research, University of Wollongong, Wollongong, Australia, and Institute for Brand Communication Research, Bergische University Wuppertal, Wuppertal, Germany

Abstract

Purpose – New measures in marketing are invariably created by using a psychometric approach based on Churchill's "scale development" procedure. This paper aims to compare and contrast Churchill's procedure with Rossiter's content-validity approach to measurement, called C-OAR-SE.

Design/methodology approach – The comparison of the two procedures is by rational argument and forms the theoretical first half of the paper. In the applied second half of the paper, three recent articles from the *Journal of Marketing* (*JM*) that introduce new constructs and measures are criticized and corrected from the C-OAR-SE perspective.

Findings – The C-OAR-SE method differs from Churchill's method by arguing for: total emphasis on achieving high content validity of the item(s) and answer scale – without which nothing else matters; use of single-item measures for "basic" constructs and for the first-order components of "abstract" constructs; abandonment of the "reflective" measurement model, along with its associated statistical techniques of factor analysis and coefficient alpha, arguing that all abstract constructs must be measured as "formative"; and abandonment of external validation methods, notably multitrait-multimethod analysis (MTMM) and structural equation modeling (SEM), to be replaced by internal content-validation of the measure itself. The C-OAR-SE method can be applied – as demonstrated in the last part of the article – by any verbally intelligent researcher. However, less confident researchers may need to seek the assistance of one or two colleagues who fully understand the new method.

Practical implications – If a measure is not highly content-valid to begin with – and none of the new measures in the *JM* articles criticized is highly content-valid – then no subsequent psychometric properties can save it. Highly content-valid measures are absolutely necessary for proper tests of theories and hypotheses, and for obtaining trustworthy findings in marketing.

Originality/value – C-OAR-SE is completely original and Rossiter's updated version should be followed. C-OAR-SE is leading the necessary marketing measurement revolution.

Keywords Construct definition, Content validity, Reliability, Marketing knowledge, Marketing, Measurement

Paper type Conceptual paper

Steven Bellman, Lars Bergkvist, Sara Dolnicar, Tobias Langner, Nick Lee, an anonymous *EJM* reviewer, and Stewart Rossiter provided very useful comments on an earlier draft of this article. Profound thanks must also go to the author's early healthily sceptical mentor in psychometrics, James Lumsden. Preparation of the article was greatly facilitated by an Australian Research Council Discovery Grant awarded to the author.

European Journal of Marketing
Vol. 45 No. 11/12, 2011
pp. 1561-1588

0309-0566
DOI 10.1108/03090561111167298

1. Introduction

Marketing knowledge, which consists of strategic principles, planning frameworks, and generalizations from empirical findings – (see Rossiter, 2001, 2002b), depends on – and indeed takes as a given – that the constructs involved have been validly measured. Valid measures are produced if researchers follow and meet the criteria spelled out in Churchill's (1979) "scale-development" procedure, which is based on Nunnally's (1978) version of psychometric theory. However, Churchill's method is dangerously misleading because it bypasses the first and fundamental requirement of the measure – content validity – and researchers following it try to "prop up" and justify low content-valid measures by claiming that the scores from these measures meet widely agreed statistical criteria. The measure is then assumed by researchers to be "valid" because it produces scores that have "good psychometric properties," all the while forgetting to ensure that the measure was content-valid to begin with. Typical examples of this cavalier (and unscientific) practice in our leading journal, the *Journal of Marketing*, are given in the second half of this article. The purpose of these critiques of psychometrically trained researchers' work is to dramatize the need for a complete "revolution" in marketing measurement.

Leading this revolution is the C-OAR-SE method (see Rossiter, 2002a, 2005, 2007, 2008, 2009a, 2011). C-OAR-SE is an acronym for its six procedural steps of Construct definition, Object representation, Attribute classification, Rater-entity identification, Scale (item type and answer format) selection, and Enumeration (scoring). C-OAR-SE is based on expert content-validation and does not use psychometrics or statistics.

The first part of the article argues in detail that the C-OAR-SE approach to measurement is incompatible with Churchill's approach and proves rationally that C-OAR-SE should be used instead. Table I provides a side-by-side comparison of Churchill's (1979) measure-development procedure and the updated C-OAR-SE procedure (see Rossiter's 2011 book – although the main updates are summarized in the present article). An explanation of the main differences between the two methods is given in the accompanying text. The C-OAR-SE measure evaluation criteria are then reviewed as a prelude to the second part of the article, which scrutinizes the definitions and measures of new constructs in recent *JM* articles from the C-OAR-SE perspective. The objective purpose of the critiques is to demonstrate how marketing knowledge is misleadingly inferred when low content-valid measures are employed. The subjective purpose, as mentioned, is to give young researchers the confidence to adopt C-OAR-SE and lead a measurement revolution.

2. Comparison of the Churchill and C-OAR-SE procedures

2.1 Different focus of the two procedures

Understanding of the major difference in focus of the two measurement procedures is helped considerably if you first look at the general structure-of-measurement model (see Figure 1). This model reveals the crucial difference in the coverage of the two procedures and also reveals the source of the problems with conventional psychometrics. Churchill's procedure (and likewise Nunnally's, 1978 procedure) covers only the "back end" (M → S) of this Construct → Measure → Score model – it attempts to validate the measure, M, by the scores, S, that it produces. In Churchill's theory of measurement, as in Nunnally's, the measure is regarded as "validated" if it yields scores that correlate highly with scores from another measure of the construct

Measurement theory and procedural steps	Churchill	C-OAR-SE
True-score theory	Based on old true-score model: Observed score = True score + Random error	Based on new true-score model: Observed score = True score + Measure-induced distortion + Rater error
Scope	Applicable only to "abstract" (multiple-item) constructs	Applies to all constructs, "concrete" (single-item) and "abstract" (multiple-item)
Validity	Content validity: Acknow-ledged as essential, but inadequately defined and handled in Churchill's measure-development procedure. Construct validity: Seen as essential, though should be called measure validity. Measure validity is wrongly tested empirically by examining convergent correlations and discriminant correlations with other measures. Predictive validity: Essential, but not adequately explained	Content validity: Essential, and consists of (a) item-content validity – semantic identity of the construct and the measure; and (b) answer-scale validity – freedom from measure-induced distortions. Established rationally by expert judgment Construct validity: Meaningless, because you cannot validate – that is, prove the truth of – a construct. You can only validate a measure of a construct, and then only by a rational argument as to its high content validity, not by any empirical means Predictive validity: Desirable but not essential. Predictive validity applies only to predictor constructs. Criterion constructs depend completely on high content validity
Reliability	Defined as absence of random (i.e. rater) error in observed scores, following the "old" true-score model. But operationalized only as internal-consistency reliability (coefficient alpha), which assumes a multiple-item measure Churchill mentions test-retest reliability (stability) but advises against using it	Stability reliability: Essential, observed score(s) must be highly repeatable on a short-interval retest Precision reliability: Accuracy of observed score(s), which depends mainly on sample size and presumes a highly content-valid measure. Precision reliability should be reported for observed scores on all the main measures in the study
1. Define the construct	Churchill defines the construct in terms of the attribute only. This mistake is made by almost all researchers	C-OAR-SE construct definition requires specification of (1) the object to be rated, (2) the attribute it is to be rated on, and (3) the rater entity, who does the rating. Constructs are ultimately researcher-defined, with no empirical assistance other than pooled experts' judgments when the researcher is unsure

(*continued*)

Table I.
comparison of Churchill's procedure and the C-OAR-SE procedure

Measurement theory and procedural steps	Churchill	C-OAR-SE
2. Generate items	Candidate items are either borrowed from others' measures (of questionable content validity and unknown stability) or are generated from qualitative open-ended interviews, with the item content mainly decided by the raters	Items must be decided on ultimately by the researcher. Raters' inputs are necessary only if the construct is perceptual. Raters' inputs are not used if the construct is psychological, i.e. not self-reportable
3. Purify the measure	Items are deleted from the candidate pool if they don't correlate with other items and with a "latent" statistical factor and don't contribute to a high coefficient alpha	Items are never deleted from the defined set of items. The items are based on a priori argued item-content validity, not derived from correlated scores ex post
4. Assess reliability	Only internal-consistency reliability (coefficient α) is calculated. Coefficient α is legitimate (though unnecessary) for a multiple-item measure but meaningless for a single-item measure. Nunnally's (1978) minimum α of 0.8 for a final measure is very often ignored and the measure is used anyway	Stability reliability is assessed by a short-interval test-retest. High stability (a "double-positive" repeat rate of .8 is the acceptable minimum) is required for the measure Precision reliability can be estimated from the sample size of raters in a particular study by using "lookup" tables
5. Assess construct validity	Construct validity is assessed by the multitrait-multimethod correlational procedure, which does not relate to the construct itself. In any case, construct validation can only mean measure validation Churchill also recommends empirically testing the measure for known-groups discriminant validity, but this is just another form of predictive validity	Constructs are definitions, not empirically testable propositions. Only a measure can be validated (with regard to the defined construct). This is content validity (high item-content validity and high answer-scale validity) and high content validity is essential Predictive validity (of the measure of a predictor construct) is desirable only, not essential. Predictive validity requires prior high content validity of the measure and a population correlation estimate against which to assess the observed predictive validity correlation
6. Develop norms	Norms are misleadingly recommended as a solution to the problem of assessing whether you're getting true scores from different answer scales. Norms require a very large and representative rater sample – rarely attained in academic studies, which usually employ college students, a nonrepresentative rater entity	Norms are needed in the form of population correlations to properly assess predictive validity Norms based on measures with low content validity, and observed-score comparisons based on a different measure than the one in the norms, are useless

Table I.

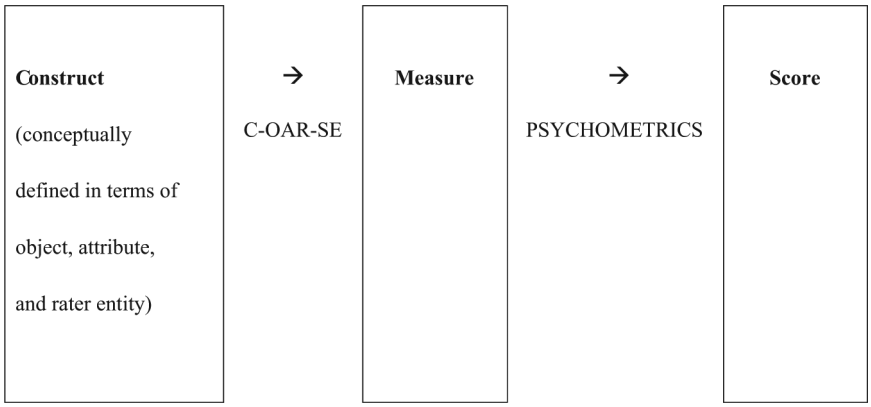

Figure 1.
General structure-of-measurement model

(convergent validity) but at the same time yields scores that do not correlate significantly with scores from a measure of another construct (discriminant validity). Churchill (1979), as does Nunnally (1978), refers to this convergent and divergent correlational procedure as establishing "construct validity." However, the construct, C, is in fact ignored, because the focus is on M and S only.

In the C-OAR-SE theory of measurement, a measure's scores are completely irrelevant to establishing the measure's validity (as too are all empirical or statistical tests). In C-OAR-SE, the validity of the measure is established solely by rational analysis – performable, for all but the deepest psychological constructs rarely used in marketing, by any verbally intelligent expert speaker of the language or, for less confident researchers or those who are not native speakers of the language in the measure, by enlisting the aid of one or two verbally intelligent expert speakers – of the semantic correspondence between the construct, C, as defined by the researcher, and the measure, M. The focus in C-OAR-SE is on C and M only. This rational analysis must demonstrate high content validity for the measure. Otherwise, the measure must not be used. What happens all too often at present with researchers who use Churchill's (or Nunnally's) procedure, as explained later in this article, is that items are added or dropped until the "alpha" is pushed up high enough to justify use of the measure. Incidentally, it is hardly ever at the $\alpha = 0.80$ minimum to 0.95 maximum as recommended by Nunnally (1978, pp. 245-246) for a "final" measure. For example, all three studies criticized later in this article used multiple-item measures with alphas below 0.80. In C-OAR-SE, the "alpha" of the measure is irrelevant. No *ex post* statistical manipulations of the items' scores can compensate for low content validity of the measure.

The focus of C-OAR-SE theory is thus on the "front end" (C → M) of the Construct → Measure → Score model. The psychometric "back end" (M → S) is immaterial and, much worse, misleading. Failure to understand the different focus of the two measurement methods has led editors and reviewers – for example at *JM* and also *JMR* – to regularly request that the present author "prove empirically" that C-OAR-SE produces more valid measures; in other words, they want the author to prove that C-OAR-SE is superior by using the very procedure that it is designed to replace! That's both philosophically (logically) impossible and it's beside the point, because Churchill's procedure is anyway fatally flawed. C-OAR-SE theory is based on

rational analysis. As with mathematics and logic, to ask for empirical proof is to ask for the unnecessary, and to demand that C-OAR-SE theory be justified by any empirical method, such as psychometrics or statistics, is to miss the purpose and value of the new procedure.

2.2 Detailed comparison

A detailed comparison of Churchill's method and the C-OAR-SE method is provided in a summary table (Table 1). The main differences are explained in the remaining sections of the first part of this article.

2.3 Measurement theory comparison

The Churchill and C-OAR-SE methods differ fundamentally in their assumptions about measurement theory. As indicated in the table, the differences pertain to their respectively underlying true-score theories, their scope of application, and the way they define and assess validity and reliability.

True-score theory. Churchill's (and also Nunnally's) measurement theory claims to be based on what may be called the "revised" true-score model, which is Observed score = True score + Systematic error + Random error (Churchill, 1979, p. 65). However, the second term in this model, "systematic error," is not clearly defined by Churchill and neither is it referred to subsequently in his article. In effect, his measurement procedure falls back on the "old," or "classical test-theory," true-score model (see Spearman, 1904; Lord and Novick, 1968) in which there is no term called "systematic error." The "old" true-score model is Observed score = True score + Random error, where "random error" (deviations from the true score) is attributed entirely to transient and presumably randomly occurring mistakes made by the rater. Churchill's methods of establishing validity and reliability (see below) depend on correlations of the scores, and the correlation statistic is based on the "old" true-score model in which there is only "random error."

C-OAR-SE, in contrast, is based on a new true-score model (also see Rossiter, 2011), which is Observed score = True score + Measure-induced distortion + Rater error. Measure-induced distortion is roughly what other true-score theorists refer to as "systematic error" (although one type of distortion due to an overdiscriminating answer scale does look "random") and it is caused by the measure, specifically by its inadequate content validity. Rater error, the final term in the new true-score model, is the same as "random error" in the old or classical true-score model, but rater error will be negligible if the measure is highly content-valid because a highly content-valid measure produces very little or ideally no distortion and raters are most unlikely to make mistaken ratings when using such a measure. The Observed score from a highly content-valid measure should therefore be the True score.

Scope of application. Churchill's theory is "only applicable to multi-item measures" (p. 66). Churchill advocates the use of multiple-item measures for all constructs, stating plainly that "marketers are much better served with multi-item than single-item measures of their constructs" (p. 66). His recommendation to always use multiple items to measure a construct – a recommendation accepted and followed by most if not all academic marketing and social science researchers – inadvertently eliminates measures of the most commonly measured construct in the social sciences and marketing: beliefs, or as marketing scientists call them, perceptions. A belief or

perception is always uncontestably measured with a single item. And indeed, all "basic" constructs – called "doubly concrete" constructs in C-OAR-SE theory – are most validly measured with a single item. This was suggested in the study by Bergkvist and Rossiter (2007) and confirmed in a reanalysis by Rossiter and Bergkvist (2009).

Even further-reaching in C-OAR-SE theory is the realization that all "complex" or "abstract" constructs – the type of construct to which Churchill's theory refers – are no more than aggregations of "doubly concrete" constructs (beliefs mostly) and these are each measured uncontestably with a single item. This point was made most clearly in the article by Rossiter and Bergkvist (2009). Take as an example the "abstract" construct COMPANY A's SERVICE QUALITY AS PERCEIVED BY ITS CUSTOMERS, as measured by the now-standard SERVQUAL questionnaire (see Parasuraman *et al.*, 1994). Each of the 21 items in SERVQUAL is a belief and therefore is a "doubly concrete" construct that requires only a single item to measure it. Item 21 in the 1994 version of SERVQUAL, for example, is the belief that "Company A has convenient business hours," and this, like the other 20, is a single-item measure. Service quality researchers then proceed – unnecessarily – to factor-analyze these single items, reducing the scores to five "factors" or "dimensions" that have no real-world applicability. How, for instance, can the marketer possibly implement the SERVQUAL factor of "Responsiveness"? Answer: Only by going back to the original single items. But worse, the marketer would not realize that important "responsiveness" items may have been omitted because they didn't "load" significantly on the "Responsiveness factor." This illustrates a major problem with Churchill's procedure, which is that, like all classic psychometric approaches, it assumes that all attributes are "reflective." The attributes of "Responsiveness," "Empathy" and so forth in SERVQUAL are clearly "formed," not reflective – formed from the most prevalent specific behaviors that make up the attribute. For "Responsiveness," for example, specific formative behaviors would include answering customers' phone calls promptly and fixing problems fully. Prevalent attribute-forming behaviors – items – cannot be discarded merely because their scores are not "unidimensional" with the other items' scores. The reflective measurement model de-validates multiple-item measures.

In their reliance on the reflective measurement model, researchers have missed or preferred to ignore Armstrong's (Armstrong, 1967; Armstrong and Soelberg, 1968) devastating demonstrations of the misleading nature of factor analysis. In Armstrong's two studies, meaningful "factors" were obtained using random numbers as inputs. Armstrong revealed the input data were random only after the "results" had been plausibly written up. Factor analysis and its principal components variation cannot be trusted (see also Ehrenberg, 1975) and this "data reduction" technique should not be used.

Churchill's procedure therefore has limited scope. The C-OAR-SE procedure has no limitation, applying to all types of construct, including the most common construct in the social sciences – beliefs or perceptions – which are *always* measured with single items.

2.4 Validity: Churchill vs. C-OAR-SE

As is conventional in psychometric theory, Churchill separates validity into content validity, construct validity, and predictive validity. In C-OAR-SE theory, only content

validity matters. Whereas Churchill concurs that high content validity of a measure is essential, he conceptualizes content validity inadequately and he unjustifiably assumes it to be achieved in steps 1 and 2 of his procedure (see shortly).

Content validity. High content validity of the measure is essential in C-OAR-SE theory and is the only requirement for a measure. What, though, does "high content validity" mean? Well, according to an informative study by Mosteller and Youtz (1990) the average person takes the adjective "high" to mean just over 80 percent probability. Quantitatively oriented readers, therefore, can take "high content validity" to mean that the semantic content of the measure must have at least 80 percent correspondence, or "overlap," with the semantic content of the construct as defined. The semantic overlap could be quantified by anyone familiar with a thesaurus and with Osgood's (see Osgood *et al.*, 1957) measure of connotative meaning. Qualitatively, however, high content validity in C-OAR-SE is simply a matter of making a rational argument – an expert-judgment appeal to readers (see Nunnally, 1978, p. 94) that the descriptor "high" is warranted for the content validity of the measure.

Content validity is defined in C-OAR-SE theory as consisting of two parts:

(1) item-content validity (which means coming as close as possible to semantic identity between the content of the construct, as defined by the researcher, and the content of the question part of the measure); and

(2) answer-scale validity (which means freedom from measure-induced distortions of the true score caused by semantic confusion when the rater is responding to the answer part of the measure).

This two-part definition of content validity is unique to C-OAR-SE.

Construct validity. So-called "construct validity" in Churchill's theory, as in Nunnally's, is seen as separate from, and apparently more important than, content validity. (Churchill states, on p. 70, that a measure that is content-valid "may or may not produce a measure which has construct validity.") Churchill, and Nunnally, advocate the usual statistical psychometric approach to establishing construct validity, which is Campbell and Fiske's (1959) multitrait-multimethod, or MTMM, analysis. But MTMM analysis focuses on the scores from measures. In MTMM, "convergent validity" and "discriminant validity" are claimed for the measure without considering the construct (see the structure-of-measurement model in Figure 1 earlier). By ignoring the measure's correspondence with the construct, MTMM analysis therefore fails to consider the measure's content validity.

Not revealed by psychometricians is that the very term "construct validity" is a logical impossibility – a misnomer. Nunnally (1978, p. 109) appears to realize the illogic but ignores it and Churchill (1979) obviously does not realize it because in his article he elevates "construct validity" over all other forms of validity. But one can never validate a construct. A construct is always just a subjective theoretical definition – a matter of opinion, not provable fact. A construct can be "reasonable or unreasonable" (in its definition) but it can't be "true or false" (validated).

Validity refers only to a measure: it is the extent to which a measure "measures what it is supposed to measure" (a definition attributed to Kelley, 1927; and see almost any social science research textbook for acknowledgement that this is the real meaning of "validity" – followed by most textbook writers' immediate departures into other "psychometric" meanings of the term!). What the measure is supposed to measure is

the construct. A measure has high validity – high truth value – only if its content closely represents the content of the defined construct.

Predictive validity. Churchill further postulates that predictive validity is important to establish for a measure. In his theory, Churchill appears to regard predictive validity as essential (see his discussion, on p. 72 of his 1979 article, of whether a measure "behaves as expected"). However, predictive validity can at most be desirable, not essential. Predictive validity cannot logically be essential because validity, by definition, is internal to the measure, and so validity cannot be established "externally" by showing that scores on the measure predict those from another measure.

Churchill also misses Nunnally's observation (Nunnally, 1978, p. 91) that predictive validity applies only to measures of predictor constructs. Measures of criterion constructs can be validated only in terms of their content validity.

Predictive validity is also much more complicated to establish than Churchill realizes, because it requires comparison of the observed correlation (called the "validity coefficient"; see Cronbach, 1961) with an estimate of the population correlation between scores on the predictor measure and scores on the criterion measure (see Rossiter, 2002a, pp. 327-328, and also see the study by Rossiter and Bergkvist (2009), in which two population correlations are estimated).

Nomological validity. So-called "nomological" or "theoretical network" validity (the main rationale for "structural equation modeling" – see Bagozzi, 1994) is just another form of predictive validity. Nomological validity, too, is merely desirable, not essential, for a measure.

2.5 Reliability: Churchill vs. C-OAR-SE

There are two principal types of reliability written about in the measurement literature:

(1) test-retest reliability, or stability, which applies to all measures; and

(2) internal-consistency reliability, which applies only to multiple-item measures.

Both forms of reliability are defined by psychometricians as the absence of "random error" (i.e. rater error) and thereby adhere to the *old* true-score model. Both forms ignore possible "systematic error" caused by the measure – that is, measure-induced distortion – which is a key term in the new true-score model presented earlier.

Test-retest reliability. Churchill states categorically that test-retest reliability "should not be used" (p. 70). This type of reliability was also dismissed by Rossiter in the initial version of C-OAR-SE (Rossiter, 2002a, p. 328). Both theorists' reason for rejecting test-retest reliability was that even a totally non-valid measure could produce highly similar scores on the retest. However, in the new version of C-OAR-SE theory (here, and in Rossiter, 2011) it is recognized that the converse does not hold. The measure must produce stable scores when readministered to the same respondents over a short retest interval, otherwise the results from any one-off empirical study using the measure cannot be trusted. Highly stable test-retest scores are guaranteed only for a measure that has high item-content validity and high answer-scale validity – that is, high overall content validity.

Internal-consistency reliability. Churchill puts the entire emphasis in his theory of reliability on internal-consistency reliability. Internal-consistency (of scores from a multiple-item measure) is invariably assessed by calculating Cronbach's (1951) coefficient alpha, symbolized α. In the original C-OAR-SE article, Rossiter (2002a)

supported the use of coefficient α – preceded by the use of Revelle's little-known (Revelle, 1979) coefficient β – in two of the six cells of the theory. These were the cells in which a "reflective attribute," which Rossiter (2002a) called an eliciting attribute, is part of the construct. The most radical update in C-OAR-SE theory (here, and see Rossiter, 2011) is to reject the "reflective" model (which means that all abstract attributes now follow the "formative" model). This change also makes internal consistency – and with it the psychometric idea of unidimensionality – unnecessary and indeed harmful. The harmful aspect is that the attainment of high internal consistency always lowers the content validity of the measure.

The new proposition in C-OAR-SE that all abstract attributes follow the formative model – that is, that the total abstract attribute score is formed from its attribute component scores – is radical given that several leading theorists (e.g. Borsboom, 2005) do not regard the "formative" approach to be legitimate "measurement" (because they cling to the unnecessary psychometric concept of "unidimensionality" – see Rossiter (2011), for a critique of this concept). The new proposition therefore requires some justification. The argument is twofold.

First, all abstract attributes (an abstract attribute has more than one clear meaning) must be classified on the basis of theory as either a formed, achieved attribute or an eliciting, dispositional attribute. The classification cannot be made empirically, contrary to the approach advocated by psychometricians such as Diamantopoulos and Sigauw (2006) and see Rossiter (2008) for a thorough dismantling of their empirical approach. The majority of abstract attributes in marketing are *formed*. MARKET ORIENTATION, SERVICE QUALITY, and CUSTOMER BRAND EQUITY are major examples of "formative" constructs that are invariably measured wrongly as "reflective" (you can tell this easily if "factor analysis" or alternatively "principal components analysis," the favored statistical tool of psychometricians, is mentioned in the measure-development section of the article). All three *JM* articles critiqued later in this article make this mistake of imposing the reflective measurement model on the measure of their new construct. Very few abstract attributes in marketing are genuinely "eliciting," or "dispositional," in that they are something internal (to the company or to the person) that causes (mental or overt) behavioral responses. CORPORATE VALUES and INDIVIDUAL PERSONALITY TRAITS are among the rare examples of dispositional attributes.

Second – and here's the more subtle argument – even though the component behaviors are caused, or "reflected out," by the dispositional attribute, they cannot be "sampled randomly" as assumed in the notion of "domain sampling" which underlies the reflective measurement model. Instead, the component behaviors must be in the measure, as items, by definition. This argument is well illustrated by carefully considering the nature of the attribute called MARKET ORIENTATION, which Narver and Slater (1990) defined as consisting of five components:

(1) customer orientation;
(2) competitor orientation;
(3) interfunctional orientation;
(4) long-term orientation; and
(5) profitability orientation.

In the initial C-OAR-SE article, Rossiter (2002a) argued that MARKET ORIENTATION is a formed attribute – being something that the COMPANY achieves. If so, all five components must be represented in the measurement items (whether the five components should be equally or differentially "weighted" is another decision – a theoretical, not an empirical, decision). However, it could alternatively be argued, in the up-front theory section of the researchers' article, that MARKET ORIENTATION is a disposition – a "company trait" if you like – that manifests itself or "reflects out" on the five sets of component behaviors. If so, again all five components must be represented in the measurement items because that is how the MARKET ORIENTATION attribute is defined conceptually. (By the way, no MARKET ORIENTATION researcher from Narver and Slater to the present has defined it this way – as a corporate disposition or "company trait.") But what is not realized by later researchers is that Narver and Slater (1990) imposed a reflective measurement model on the scores from their original large list of conforming to the five components items and dropped two of the components, Long-term orientation and Profit orientation, from the final measure because neither resulted in an "internally consistent" (by coefficient alpha) "factor." Their actual MARKET ORIENTATION measure represents only three of the five defined components:

- market orientation;
- competitor orientation; and
- interfunctional orientation.

They and all subsequent researchers using Narver and Slater's (1990) 15-item MARKET ORIENTATION scale are thus using a measure that does not correspond semantically with the construct definition. Adoption of the C-OAR-SE method by later researchers (it was published in 2002, well after Narver and Slater's (1990) article) would have prevented this major content omission because C-OAR-SE, as made clear by the C → M → S model earlier, is all about content validity, which requires high semantic correspondence between the construct and the measure. No matter whether the MARKET ORIENTATION attribute was conceptualized as "formed" or "reflective," all five components would be properly represented in the measure if the C-OAR-SE procedure were followed.

When you think about it, therefore, all abstract attributes are formed from a measurement standpoint (formed from their predefined components). The radical "fallout" from careful thought is that the reflective measurement model is entirely misleading and should be abandoned.

Precision reliability. C-OAR-SE theory adds another form of reliability much valued by social science practitioners: precision-of-score reliability (abbreviated as precision reliability in the comparison table). Churchill (1979, p. 66) hints at this form of reliability when he dismisses single-item measures. His argument is that the usual seven-step rating scale accompanying a single-item measure produces imprecise scores because "the same scale position is unlikely to be checked in successive administrations". This may be true but this is test-retest (un)reliability, not precision reliability.

In C-OAR-SE theory, precision reliability is much closer to what practitioners know to be a very important applied consideration, which is the confidence that can be placed in an observed score. "Precision" in this particular meaning of reliability depends

mainly on the sample size of observations and therefore can be estimated closely enough for practical purposes from "lookup" tables which give generalized 95 percent confidence intervals for various sample sizes (see Rossiter and Percy, 1987; Rossiter and Percy, 1997; Rossiter and Bellman, 2005; and Rossiter, 2011). These lookup tables are what opinion pollsters use in newspaper reports and sometimes in TV reports to predict elections and if the precision in lookup tables is accurate enough for measuring important societal and political knowledge it is surely good enough for assessing the incidence of marketing knowledge.

All precision estimates depend on using content-valid measures in the first place. If the measures are highly content-valid, then the statistical precision of scores derived from them becomes relevant for proper interpretation of the findings.

2.6 Step-by-step comparison

A brief comparison of the other major differences between the C-OAR-SE procedure and the six steps in Churchill's (1979) measure-development procedure concludes the first half of this article. The comparison is made on the basis of Churchill's six steps.

1. Define the construct. Churchill – as all other psychometricians do – defines the construct in terms of its attribute only. Churchill encourages this when he makes the incorrect (from the C-OAR-SE perspective) comment on p. 65 of his article that ". . .it is the attributes of objects that are measured and not the objects themselves." In C-OAR-SE, the construct must be defined in terms of the object to be rated, the attribute it is to be rated on, and the rater entity doing the rating.

McGuire (1989), in his "object-on-attribute" conceptualization of constructs, explains why a construct is necessarily "underspecified" if the object is not included in the construct definition. It follows that the measure, also, must represent the object of the construct – for example, the measure must include an illustration of the product if such products are usually chosen by brand recognition, or must include the phonetically appropriate name of the product or service if it is usually chosen by brand recall – see Rossiter and Percy (1987, 1997) or Rossiter and Bellman (2005). Object misrepresentation is one of the most common measurement mistakes made by researchers. It is a mistake of low item-content validity.

The attribute is only the second element of the construct. It, too, must be correctly represented in the measure. Churchill's assumption that all attributes have multiple meanings and therefore are "multidimensional" automatically excludes from his theory all single-meaning attributes. Single-meaning attributes are the type of attribute represented in the most common construct in the social sciences – beliefs or perceptions.

The third element of the construct, the rater entity, does not appear in the measure, but must be included in the definition. For example, service quality researchers should define as two constructs MANAGER-RATED service quality of the organization, on the one hand, and CUSTOMER-RATED service quality of the organization, on the other. These two constructs represent the main service-quality "gap" that marketers must manage. Most meta-analyses fail to identify the various different rater entities (see Rossiter (2011), for examples of this). Rater-entity differences are a major reason for reaching the unsatisfactory conclusion of "mixed" findings.

2. Generate items. The second step in Churchill's procedure is item generation, which always means the generation of multiple items. In Churchill's procedure,

candidate items are either generated from qualitative "open-ended" interviews with a sample of raters, or else far more often they are borrowed from other researchers' measures. The second of these methods of item generation can now be readily seen as flawed given the near certainty that previous measures will have questionable content validity (low item-content validity) as well as unknown stability or test-retest reliability (low stability is largely caused by low answer-scale validity, which leads raters to mark the scale differently each time).

The former method of generating candidate items by conducting open-ended interviews with a sample of raters is the "textbook correct" method (it's correct even for a well-established object and attribute, because the rater entity might be different). However, this method is inappropriate for generating an item or items for a "psychological" construct – defined in Rossiter (2011) as a construct that is not self-reportable by raters (examples in psychology would be the Freudian constructs of REPRESSION and PROJECTION; the increasingly popular construct in psychology and also in consumer behavior of IMPLICIT ATTITUDE; and the very important and inadequately measured set of constructs in both disciplines known as MOTIVES, noting that qualitative research was originally called "motivation research"). The item or items used in the measure of a psychological construct can be decided only by the researcher, and raters are of no help.

In fact, the final item, or items, selected for the other type of construct – called a "perceptual" construct in Rossiter (2011) because it *is* self-reportable by the rater – must also be decided on ultimately by the researcher, although pretesting of item-wording with raters is a good idea if the researcher is unsure of "consumer language" terms for the attributes.

3. Purify the measure. Churchill's notion of "purifying" the (multiple-item) measure is a nice-sounding but misleading religious metaphor. In the "purification" step, items are deleted from the randomly generated pool of candidate items if their scores fail to correlate positively with each other and with the total score on a "latent" and entirely artifactual, statistically derived "factor" emerging from the usually performed factor analysis or principal components analysis of candidate items' scores. The fact that an object receives high scores on, say, the item "Likable" and the item "Honest" (i.e. their scores are highly correlated) does not mean that there exists a real attribute labeled "LIKABILITY/HONESTY" – yet this is what a factor analyst will infer! The five SERVQUAL "dimensions" of Responsiveness, Empathy, etc., are typical examples of the factor-analysis fallacy. To make matters worse, further items may be deleted if the high-loading items fail to produce a high coefficient alpha. This "purification" step is really a "contamination" step, because a multiple-item measure with poorer content validity is always the result when defining items are deleted or when their scores are summarized as an artificial "factor."

There is no "purification" step in C-OAR-SE. An abstract object or an abstract attribute means, of course, that a multiple-item measure must be employed, but the multiple items are "in there" by definition, having previously been selected – ultimately by the researcher – as corresponding to the components in the definition. Each item is based on prior certification by the researcher that the item has high item-content validity, that is, high semantic correspondence with its predefined component in the researcher's construct definition. Scores from the items are never to

be considered in assessing the validity of the measure (see the C → M → S model earlier) and thus no "purification" step is needed.

4. Assess reliability. As pointed out earlier in this article, "reliability" in Churchill's theory refers solely to internal-consistency reliability (as estimated by calculating coefficient α from the items' scores) and applies only to a multiple-item measure. But, in C-OAR-SE theory, internal consistency is irrelevant and misleading for (scores on) a multiple-item measure because the total score on such a measure is always formed from the scores obtained on the items measuring the predefined components and these scores do not need to be internally consistent or at all correlated (although they usually will be, given that the attribute components are components of the main attribute).

In C-OAR-SE, there are only two types of reliability that matter. These are stability reliability and precision reliability. They were explained in the section on "reliability" earlier.

5. Assess construct validity. The first sub-heading under step 5 in Churchill's (1979) article, the step describing "construct validity," is "Correlations With Other Measures." In this section, Churchill goes into great detail to exemplify how the correlational theory of construct validity known as multitrait-multimethod analysis, or MTMM, an analysis procedure invented by Campbell and Fiske (1959), is to be applied to the measure. However, a founding principle of C-OAR-SE theory, represented in the C → M → S model (see earlier figure), is that a measure cannot be validated in relation to a construct by examining the scores obtained from the measure (in the form of the scores' convergent, discriminant, or predictive correlations, coefficient alpha, or any other statistic). In C-OAR-SE theory, the concept of "construct validity" is replaced by content validity.

Content validity requires a rational argument – made by the researcher as theorist if a "psychological" construct and as an expert in colloquial consumer language if a "perceptual" construct, aided if necessary in either case by a couple of expert colleagues – that there is very good semantic correspondence between the construct as defined and the measure as selected (item-content validity) plus certification by the researcher, perhaps aided by a pretest with a small sample of raters, that the answer scale selected for the measure has very good "expressability" (high answer-scale validity). C-OAR-SE measurement items require only a rational supporting argument attesting that they are highly content-valid. Most researchers are evidently capable of doing this content analysis on their own. Hardesty and Bearden (2004), for instance, estimate that multiple expert judges were used for only about 20 percent of the approximately 200 new measures reported in Bearden and Netemeyer's (1999) handbook of marketing measures – that is, in about 80 percent of cases, the researcher alone designed the new measure.

6. Develop norms. The final step in Churchill's measure-development procedure is to "develop norms" for scores obtained from various applications of the measure. As he points out (on p. 72), this final step is necessary only if the researcher wants to compare the scores of individuals – or the scores of individual objects, such as a company, a brand, an ad, or celebrity or politician – with some population average score (the "norm"). But few studies in the social sciences have this purpose and so it is not a necessary step. Norms do have their uses, however. Psychologists often use norms in research on individual abilities; well-known examples include the testing of GENERAL MENTAL ABILITY (called GENERAL INTELLIGENCE or "IQ" before the political

correctness movement descended upon us) and the measuring of individuals' psychological PERSONALITY TRAITS, which are assessed for the "clinical" population relative to their average levels in the "normal" population. Marketing practitioners sometimes use norms when they use marketing models (e.g. the BASS DIFFUSION MODEL, and the ORDER-OF-ENTRY → MARKET SHARE MODEL; see Urban and Star, 1991, and also Rossiter and Percy, 1987, 1997).

The greatest need for normative estimates for measures in the social sciences has been surprisingly overlooked. Normative (i.e. population-based) correlation coefficients are needed for assessing predictive validity, because good predictive validity means coming close to the true correlation, not searching statistically for the highest correlation. Only a few researchers in psychology have attempted to estimate population correlation coefficients. Important attempts are for the correlation between ATTITUDE and subsequent BEHAVIOR (Krauss, 1995) and for HABIT and INTENTION as dual predictors of BEHAVIOR (Ouellette and Wood, 1998). In marketing, Rossiter and Berkgvist (2009) have attempted to estimate the true population correlation for AD LIKING predicting BRAND ATTITUDE and then BRAND ATTITUDE predicting BRAND PURCHASE INTENTION. All estimates of population correlations rely on meta-analyses – and on the ability of the researcher to correct for problems with meta-analyses, of which differing measures are the main problem (another big problem is that college students are often the only rater entity, a problem pointed out long ago by Peterson *et al.*, 1985). The measure-difference problem would be solved if all social science researchers adopted the C-OAR-SE measurement procedure. The rater-entity difference problem can only be solved by judiciously seeking out practitioner studies based on broader populations of respondents (see Rossiter and Percy, 1987, 1997, for numerous examples).

3. C-OAR-SE critique of three *JM* studies

In this last part of the article, the main defined construct and measure in three recent articles selected from the *Journal of Marketing*, the most prestigious journal in our field, are critiqued from a C-OAR-SE perspective (See Table II). Most of the researchers involved are very experienced and have previous publications in *JM* and in other leading marketing research journals. The critiques do not criticize the researchers – except obliquely for failing to adopt the C-OAR-SE procedure, which was published well before these studies were conducted. The sole purpose of the critiques is to make readers realize that much of the marketing knowledge in strategic principles and empirical generalizations derived from studies using what Leeflang *et al.* (2009) called "soft data" social-science constructs is at the very least questionable.

The critiques are organized in terms of five C-OAR-SE-based criteria that should be discernible from the first part of the present article. (These five criteria provide a useful summary of the C-OAR-SE steps – that is, the steps that should be followed in designing or choosing a measure.) The criteria are:

(1) comprehensive conceptual definition of the construct in terms of object, attribute, and rater entity;

(2) close semantic correspondence of measurement item, or items, with the construct as defined (high item-content validity);

Table II.
C-OAR-SE critique of measures of the main constructs in three typical *JM* articles

C-OAR-SE criterion	"Brand experience" (Brakus *et al.*, JM, May 2009)	"Customer need knowledge" (Homburg *et al.*, JM, July 2009)	"Corporate culture" (Tellis *et al.*, JM, January 2009)
1. Adequate conceptual definition of the construct in terms of object, attribute, and rater entity	Yes. Though their definition did not clearly specify the rater entity	No. Construct's attribute is more accurately labeled "customer-need perception"	No. Construct's object should be defined as "organizational values"
2. Close semantic correspondence of measurement item(s) with the construct as defined (high item-content validity)	No. The items' objects and attributes completely miss the component objects and component attributes of the construct	No. Attribute-content of items too narrow and not representative of the component attributes in the construct. Task instruction for raters ambiguous with regard to the construct definition	No. Items selected so as to guarantee a high correlation between the predictor measure's scores and the criterion measure's scores (i.e. to circularly prove the researchers' theory)
3. Good "expressability" of the answer options (high answer-scale validity)	No. The seven-point unipolar answer scale probably "overdiscriminates"	No. Rank-ordering of attributes undoubtedly both "underdiscriminates" absolute differences and "overdis- criminates" likely tied ranks	No. The bipolar Likert answer scales are faulty on the "disagree" side and "reversed" items cause rater errors
4. All major defining items retained in the measure	No. Valid defining items in qualitative research are lost in moving to the quantitative measure	No. The items are extracted from an unreported set of items from qualitative research in which two of three rater entities are not valid	No. Factor analysis and coefficient alpha are used to wrongly delete defining items
5. Correct scoring rule applied to the scores	Yes. Sum-score rule (but a simple frequency count could have been used with the qualitative measure replacing the quantitative measure)	Yes. Rater-entity difference rule (but its computation is wrongly described)	Yes. Sum-score rule (but the components received unequal weight due to differing numbers of items per component)

(3) good "expressability" of the answer options (high answer-scale validity);
(4) all major defining items retained in the measure; and
(5) correct scoring rule applied to the scores.

These five criteria are hierarchical. In decision-theory terms, they form an "elimination-by-aspects" decision rule, based on C-OAR-SE, for accepting the measure as valid – or, of course, for rejecting it.

3.1 "Brand experience" (Brakus et al.*, 2009)*

In their *JM* article of May 2009, researchers Brakus, Schmitt, and Zarantonello set out to measure a new construct that they called "brand experience."

1. Adequate conceptual definition of the construct. Brakus *et al.* (2009, p. 52) defined the construct of BRAND EXPERIENCE as "sensations, feelings, cognitions, and behavioral responses evoked by brand-related stimuli that are part of a brand's design and identity, packaging, communications, and environments". In terms of C-OAR-SE theory, this is a mostly adequate conceptual definition because it specifies the object's components (the Brand, its Packaging, its Communications, and its Retail environment) and also the components of the abstract attribute of brand experience (Sensations, Feelings, Cognitions, and Behavioral responses). Their definition fails only to specify the rater entity in the construct (which can be inferred to be ALL CONSUMERS AWARE OF THE BRAND, whether or not they are customers of the brand).

2. High item-content validity. The most serious problem occurs in the researchers' measure of the construct. Without realizing that they had done so, the researchers developed the most valid measure of the construct of a "brand experience" (most valid according to C-OAR-SE) in their pre-study. Examples of BRAND EXPERIENCES (plural, note) obtained in open-ended questioning for some of the brands they studied are reproduced in Figure 2 (from their Table I, on p. 56). These verbatim self-reports clearly are brand experiences.

However, these "sensations," "feelings," "cognitions," and "behavioral responses" were measured in relation to the Brand-name only. This severely biases the object component in the measure because it omits the other object components from the construct as defined by the researchers, which were the brand's Packaging, Communications, and Retail environments (although the latter was obviously the object referred to in the open-ended question about Starbucks – see Figure 2).

The rater entity for the pre-study measure was also biased. The raters should have been ALL CONSUMERS WHO HAVE HEARD OF THE BRAND. Instead, the researchers interviewed only consumers who *use* the brand and therefore are more likely to have "brand experiences" to report.

But the instrument that the researchers developed to measure BRAND EXPERIENCE for the main study bears no resemblance to the construct as defined. What the researchers did was to generate items that do not measure consumers' actual sensations, etc., elicited by the brand (and by its packaging, communications, and retail environment) but instead measure consumers' vague assertions that they had such experiences in general. The 12 completely general items making up the researchers' BRAND EXPERIENCE measure (see their Table II, p. 58) are reproduced in Figure 3

Apple iPod

- I love the touch and feel of the product
- I am part of a "smarter" community
- I exercise more because of the iPod

Nike

- Makes me feel powerful
- I feel inspired to start working out
- I feel like an athlete
- The store incites me to act – put on the shoes, swing a bat

BMW

- I feel young
- I feel stylish
- It's just great to drive
- The symbol of my success

Starbucks

- Smells nice
- Visually warm
- Puts me in a better mood
- It's like being around a Barnes & Noble crowd

MasterCard

- Makes me think about the precious things in life
- I feel more youthful than when using American Express or Visa

Source: Brakus *et al.* (2009)

Figure 2.
Some specific brand experiences obtained open-end in the pre-study

for the reader's perusal. Compare the completely general item content of the items in Figure 3 with the specific contents of the open-ended reports of BRAND EXPERIENCES summarized in Figure 2 earlier. Can you see the problem? If not, try answering 12 BRAND EXPERIENCE questions yourself for, say, the Nike brand ("yes" or "no" will do for answers) and then compare those answers with the example answers for Nike in the previous table. The 12 items have zero content overlap with the construct, which was defined as specific experiences. This is the "fatal flaw" in their study. The main-study measure cannot be made acceptable by appealing, as the researchers did, to the "good statistics" it produces.

3. Summary evaluation. Brackus *et al.* (2009) did not in fact discover a new construct called "brand experience." They merely created an artificial general name for specific experiences in the form of consumers' beliefs and associations to brands – constructs that have been studied many times before.

1. This brand makes a strong impression on my visual or other senses.
2. I find this brand interesting in a sensory way.
3. This brand does not appeal to my senses.
4. This brand induces feelings and sentiments.
5. I do not have strong emotions for this brand.
6. This brand is an emotional brand.
7. I engage in physical actions and behaviors when I use this brand.
8. This brand results in bodily experiences.
9. This brand is not action oriented.
10. I engage in a lot of thinking when I encounter this brand.
11. This brand does not make me think.
12. This brand stimulates my curiosity and problem solving.

Note: The answer scale was 1 = "Not at all descriptive" to 7 = "Extremely descriptive"
Source: Brakus *et al.* (2009), Main Study

Figure 3.
The 12 completely general and impossibly vague items in the "brand experience" measure

3.2 "Customer need knowledge" (Homburg et al.*, 2009)*

In their *JM* article of July 2009, researchers Homburg, Wieseke, and Bornemann introduced a new construct that they called "customer need knowledge" and proposed a new measure of it.

1. Adequate conceptual definition of the construct. Homburg *et al.* (2009, p. 65) defined the new construct of CUSTOMER NEED KNOWLEDGE as "the extent to which a frontline employee can correctly identify a given customer's hierarchy of needs". However, the theoretical background they supplied leading up to their construct definition concerns "the accuracy of interpersonal perception" and it is semantically inaccurate to label the attribute in this construct as involving knowledge. A more appropriately descriptive label would be "Frontline employees' accuracy of perceiving the customer's needs," and a shorter attribute-only label would be CUSTOMER-NEED *PERCEPTION* (restoring the clarifying hyphen so often omitted today).

2. High item-content validity. The "fatal flaw" in the researchers' *measure* of CUSTOMER-NEED KNOWLEDGE occurs in terms of the next C-OAR-SE criterion: high item-content validity. The researchers asked a sample of (travel agency) frontline employees, as well as each employee's last customer, to rank-order six "needs" (reproduced verbatim in Figure 4). The "need" items have low content validity. To begin with, the attributes are far too vague and general (especially "Brand,"

1. Convenience (e.g., to have the least possible effort)
2. Price (e.g., to book travels with the best prices)
3. Service (e.g., the intensive consulting service by a travel agent)
4. Brand (e.g., to book brands of well-known travel companies)
5. Security (e.g., to have the security that the booked travel meets the expectations)
6. Shopping enjoyment (e.g., to have a travel planning that is fun and raises pleasant anticipation)

Note: Answer method was ranking

Source: Homburg *et al.* (2009)

Figure 4.
The six "needs" – listed verbatim here with their contradictory item wording – in the study of "customer need knowledge"

"Convenience," and "Price") and then they are contrarily made too specific (and therefore unrepresentative) by their accompanying examples. The example parenthesized in each item is actually a component attribute, and a full set of them should have been written as separate items. For instance, the attribute called CONVENIENCE in any thorough study of services – and especially in practitioners' studies – is always broken out into its components of Location convenience (for personal visits), Opening-hours convenience (for personal visits and telephone contact), and Perceived waiting time. Measurement of these components requires three separate items, not one item as these researchers used.

Each customer was asked to rank the needs in order of "importan[ce] for you with respect to travel booking" and then the employee who had served that customer was asked to rank the same needs in order of their "importan[ce] for this customer" (p. 78). The content-validity problem here is the ambiguity of the employee's task (the wording of the task instruction). It could be argued that the employee was *not* asked to estimate or "perceive" the customer's needs but rather to judge what those needs should be ("importance for this customer"). The employee's task instruction does not unambiguously lead to a measure of the accuracy of employees' perceptions of the customer's needs and so the measure does not correspond semantically well enough to the construct as defined and will produce misleading results. This mistake could have been avoided by pretesting of the instructions for the measure.

The researchers' use of low content-valid items (together with their use of ranked rather than rated items, discussed below) was undoubtedly responsible for their surprisingly weak findings with regard to the main construct. Customer-need satisfaction is the strategic principle underlying the "marketing concept" (see any marketing textbook) and yet these researchers found that the salesperson's CUSTOMER-NEED KNOWLEDGE was only weakly correlated with the customer's rated SATISFACTION with the visit and, even more practically important, only weakly correlated with the customer's rated WILLINGNESS TO PAY – that is, to pay a higher price for holiday tour packages booked with this travel agent (although the latter measure was half not-valid because two of the four items sought a customer's

willingness to pay a higher price for airfares, which is ridiculous to expect just because a particular travel agency booked the air travel). The average correlation (i.e. the predictive-validity coefficient) for the employee's CUSTOMER-NEED KNOWLEDGE predicting the customer's rated SATISFACTION was just $r = 0.16$, and for CUSTOMER-NEED KNOWLEDGE predicting WILLINGNESS TO PAY was also just $r = 0.16$. While such correlations are statistically significant, they translate practically to very small effect sizes (Cohen (1977), regards as a "small" effect size a correlation of between $r = 0.10$and$r = 0.29$). A practically minded marketing manager would likely conclude from these results that it is hardly worth training frontline employees to try to detect and fulfil "customer needs." This is not the conclusion the researchers intended, but their findings point to it.

3. High answer-scale validity. The third C-OAR-SE criterion requires good "expressability" of the answer options (see especially Viswanathan *et al.*, 2004). The researchers' measure of CUSTOMER-NEED KNOWLEDGE fails on this criterion as well. Employees and their customers were asked to rank the "needs" from 1 down to 6. But not only does the ranking procedure fail to indicate whether any of the needs were absolutely important, it also precludes the likely answer that several of the needs are equally important. The ranking method therefore both "underdiscriminates" by not using absolute ratings and "overdiscriminates" by forcing apart what could be tied ranks. With either problem, the answer method – forced ordinal ranks – has unacceptably low content validity.

4. All major defining items retained. Were the main defining items included in the CUSTOMER-NEED KNOWLEDGE measure (the fourth C-OAR-SE criterion)? Again the answer is "no." The researchers obtained the initial set of items from qualitative interviews with three rater entities – MANAGERS, EMPLOYEES, and CUSTOMERS – but only the last rater entity, CUSTOMERS, was relevant. (The purpose of the measure was to gauge employees' accuracy in perceiving their customers' needs.) The gathering of "customer needs regarding travel agency services" (p. 59) as nominated by travel agency MANAGERS and by travel agency EMPLOYEES goes outside the construct definition. The final list of six customer needs (see above) therefore cannot be guaranteed to include only the main defining items.

5. Correct scoring rule. The researchers chose what might be called a rank-difference scoring rule to derive the employees' accuracy scores (i.e. their "customer- need knowledge" scores). The validity of ranking was questioned above but, this aside, the actual *scoring rule* the researchers employed was appropriate. However, the researchers described its computation incorrectly (on p. 70) as "the sum of the absolute differences between customer and employee rankings multiplied by -1." The maximum of the absolute differences between the two rank orders of six objects is 18 and the minimum is 0 (and the midpoint, which might indicate 50 percent accuracy on the part of the employee, is 9). These scores should be reversed (not "multiplied by -1") so that a score of 18 indicates maximum accuracy and a score of 0 indicates complete inaccuracy. The positive numbers in their Table I on p. 70 (mean scores of 7.8 and 8.5 observed in their two studies) suggest that they did in fact use reversal scoring despite the wrongly reported formula, but this mistake may not be picked up by researchers attempting to replicate the study.

More serious is that the mean scores of below and just under 9 suggest that the average employee's perceptual accuracy in gauging the customer's needs did not reach

50 percent! This disappointing result may be due to the low validity of the task requested of employees who, as noted earlier, quite possibly estimated the needs that the customer should have rather than does have, and also may be due to the forced nature of the ranking procedure.

6. Summary evaluation. Contrary to the title of Homburg *et al.*'s article, their empirical findings lend very little support to the idea that managers should implement "the marketing concept." The findings seem to undermine the founding principle of marketing! If uncritically accepted by readers of *JM*, they would result in the false marketing knowledge that accurate detection of the customer's needs is of little importance.

3.3 "Corporate culture" (Tellis et al.*, 2009)*

In their *JM* article of January 2009, researchers Tellis, Prabhu, and Chandy studied the emergence of radical innovations across 17 of the world's major economies, using as their main predictor a construct they called "corporate culture."

1. Adequate conceptual definition of the construct. Tellis *et al.* (2009, p. 6) defined CORPORATE CULTURE as "a core set of attitudes and practices that are shared by members of the firm". Whereas this definition is admirable from a C-OAR-SE-theoretical perspective because it specifies the object of the construct (the FIRM) and the rater entity (MEMBERS OF THE FIRM), the definition includes a questionable conceptualization of the corporate culture attribute. Surely, CORPORATE CULTURE refers to *UNIVERSAL MANAGERIAL VALUES* subscribed to by the particular *ORGANIZATION* (as defined, for example, in *JM* by Deshpandé and Webster (1989)). The fact that others have used a similarly loose and unacceptable definition of the "corporate culture" attribute (other researchers are cited on p. 70) does not justify its adoption here. The scientifically unacceptable practice of justifying definitions – and measures – by an appeal to precedence is all too common in the social sciences and especially in marketing.

The object of the construct of CORPORATE CULTURE was defined as FIRMS IN GENERAL whereas the researchers confined their sample to MANUFACTURING FIRMS. They selected manufacturing firms presumably to give a higher chance of locating RADICAL INNOVATIONS – the product-based dependent variable in their study – and thereby excluded SERVICE FIRMS, which make up a majority of companies in some of the economies studied. Also, the rater entity in the construct was defined as MEMBERS OF THE FIRM whereas the researchers interviewed only "the vice-president for innovation or technology or the equivalent" (p. 9). This particular rater entity is hardly a representative "member of the firm"! The technology V-P would likely give a favorably biased report of the dependent variable.

2. High item-content validity. The "fatal flaw" in the measure, however, lies in the low content validity of the components selected to represent the abstract attribute of CORPORATE CULTURE. In a patently circular manner, the researchers confined their measure to attitudes and practices that tapped only the firm's orientation toward the single value of INNOVATION, which overlaps greatly with the dependent variable that the researchers were trying to predict (see Table III). The attitude components in their measure of CORPORATE CULTURE were Willingness to cannibalize, Future market orientation, and Risk tolerance, and the behavioral components were Encouragement of product champions, Incentives through innovation, and a third

Corporate culture (predictor variable)	Radical innovation (dependent variable)
A. Attitudes	Primary-measure items
1. Willingness to cannibalize (3 items)	1. "Our firm rarely introduces products that are radically different from existing products in the industry" (reverse-scored)
2. Future market focus (4 items)	2. "Our firm lags behind others in introducing products based on radically new technologies" (reverse-scored)
3. Risk tolerance (4 items)	3. "We have no difficulty in introducing new products that are radically different from existing products in the industry"
B. Practices	
4. Product champions (2 items)	
5. Incentives for enterprise (2 items)	
6a. Autonomy (2 items)	
6b. Internal competition (2 items)	

Note: All items answered on bipolar Likert answer scales (wrongly scored unipolar as 1 = "Strongly disagree," through 7 = "Strongly agree")
Source: Tellis *et al.* (2009)

Table III.
The six unacceptably narrow components in the measure of "corporate culture" and the subjective and redundant items used to measure the dependent variable of "radical innovation"

vaguely labeled and mixed component that they labeled as Internal markets. All the items in their CORPORATE CULTURE measure pertain narrowly to innovation. Now look at the three items they used to measure the dependent variable of RADICAL INNOVATION (also given in Table III). There is a high degree of overlap between the measures of predictor variable and the dependent variable, which renders their "theory" predictively circular.

What is simply not credible therefore, given this circularity, is that the scores on the six measured components of CORPORATE CULTURE had such small correlations with the scores on the dependent variable of what should have been labeled RADICAL INNOVATION *ATTITUDE* (this variable was subjective; the number of actual radically innovative products produced by the firm should have been used as the dependent variable). The largest correlation was $r = 0.25$, for Risk tolerance (see their Figure 3 on p. 14), and the correlations for the other five also favorably biased components ranged from 0.11 down to statistically zero ($r = 0.06, n.s.$). The zero correlation was for the firm's having Internal markets, a component that was mis-measured – see the two sets of items, one labeled "Autonomy," which has nothing to do with "Internal markets," and the other labeled "Internal competition," which does. These results are hardly convincing evidence for even the limited theory that the researchers tested.

What the researchers should have done to achieve a highly content-valid measure of CORPORATE CULTURE – more correctly labeled as ORGANIZATIONAL VALUES – was, ideally, to have generated a new C-OAR-SE-based measure by qualitatively interviewing a cross-section of top managers. For topic areas, they could use the comprehensive review of organizational behavior by Gelfand *et al.* (2007). Alternatively, they could have borrowed a more comprehensive extant measure. The four-component, constant sum measure reported in *JM* by Deshpandé *et al.* (1993)

includes as only one option in one component the item "Commitment to innovation and development." With Deshpandé *et al.*'s measure – which is closer to what C-OAR-SE would suggest – the researchers would have avoided the narrowness of their study and its patent circularity.

3. The other C-OAR-SE criteria. The other mistakes of measurement made by Tellis *et al.* (2009) are not trivial, although they pale in comparison with the major mistakes identified above. There is the error of low answer-scale validity with the Likert answer scales (see Rossiter, 2002a) used for all items for all the constructs, and so common-methods bias may have inflated the already weak correlations. There is the error of omitting defining items in the predictor measure of CORPORATE CULTURE, omissions made likely by assuming a reflective measurement model. Finally, there are unjustified unequal weights of the components in the sum-scoring of the predictor measure due to the differing number of items used per component.

4. Summary evaluation. Tellis *et al.* (2009), in their article did not contribute any new marketing knowledge. Their study did not employ a valid measure of the construct of "corporate culture" nor a valid measure of "radical innovation," and accordingly they recorded implausibly weak – and untrustable – results. The lack of a contribution was due to inadequate construct definition and poor selection of measures – mistakes that would not have occurred had the researchers followed the C-OAR-SE procedure.

3.4 Summary statement regarding the selection of the three JM articles and the generality of the critique

It is necessary to reemphasize that these three articles were purposefully but representatively selected. The three articles were purposefully selected because each introduced a potentially important new construct – BRAND EXPERIENCE, CUSTOMER-NEED PERCEPTION, and what might be reconceptualized as CORPORATE *INNOVATION* CULTURE – together with a new measure of each that was intended to contribute new marketing knowledge. It cannot reasonably be contended that these were unrepresentative, atypical articles. All passed expert review and were published in recent issues of our leading journal.

To further dispel the objection that I have been selective and that the majority of our measures in marketing are "okay," I add that all other empirical articles based on "soft" measures of established marketing constructs post-dating Churchill's (1979) article could be similarly criticized, as could similar articles in all other leading social science journals (those articles which post-date Nunnally's highly influential 1978 book on psychometric theory). I criticize – and correct – measures published in a broad range of marketing, management, organizational behavior, psychology, and sociology journals in my new book on C-OAR-SE.

This is no "straw man" critique of current social science measures. The "psychometric" measurement problem is pandemic, and a cure is urgently needed. The only cure is to adopt the C-OAR-SE method.

4. Conclusions

Social science researchers must get much braver – they must trust their ability to define new constructs (and properly define old ones) and to design highly content-valid measures of them. Amazing to many, all that is required is expertise in the colloquial language – the semantics – to be used in the measure. To give a typical and very

topical example: Is it still appropriate to describe a person – or a brand – as "cool"? Can the same meaning possibly be captured by the "hot-cold" measure in a battery of semantic-differential items? And what is the semantic opposite of "cool"? It's certainly not "hot," which has an entirely different meaning when applied to a person today and yet another meaning when applied to a brand as the object. Researchers – especially academic researchers – fail to recognize the fundamentally semantic nature of measurement. A Churchill-inspired researcher would likely borrow or invent loose multiple items representing a vague and usually undefined "domain," put them in a questionnaire with faulty Likert answer scales, show that after deleting some items the scores on the remaining items correlate and produce a "high alpha," and then claim to have captured the essence of "coolness"! This is exactly what Churchill's approach would tell the researcher to do – and the researcher is much more likely to have the work published by following it. The C-OAR-SE researcher, in contrast, would pick up these "soft" attributes during qualitative research (see Rossiter, 2009b, 2011). The C-OAR-SE researcher would realize that a literal single item (e.g. "Brand X is cool. □ Yes □ No") is perfectly valid for the particular rater entity identified in the construct definition, then bravely argue for this measure and use it. Hopefully the brave researcher will have become confident that this is the right approach by reading and understanding the present article.

Also, what is not required for valid measurement of social science constructs is expertise in the substantive field. Substantive expertise does not guarantee better measures. The three *JM* studies criticized in the present article reveal this only too well. Nor is expertise needed in quantitative methods and statistics – indeed, such expertise might be considered a liability given that C-OAR-SE is a nonstatistical theory.

This last point is important to emphasize. Measurement of "soft" (i.e. social science) constructs has been plagued by misplaced reliance on statistics – and especially psychometrics. It is logically impossible for statistical manipulations to substitute for the fundamentally conceptual task of defining constructs and the semantic task of devising valid measures of them. Yet all social sciences, including marketing, are being taken over by "scientism" (an exaggerated belief in the power of scientist-invented techniques, most of all statistical techniques). One has only to peruse the journals of today to see this, especially if the articles are compared with articles on the same constructs written before the psychometricians took over. The *Journal of Marketing*, from which the criticized studies were taken, is no exception. The best articles on "soft" constructs were written in *JM* by theorists such as Alderson, Bartels, Converse, Hunt, Kotler, Levitt, Levy, Stainton, Webster, Wensley, and Zaltman – without resort to statistics.

Social science knowledge, and therefore much of our marketing knowledge, is based on the presumption – and it *is* merely a presumption – that the measures of the many "soft" constructs involved in the knowledge are highly valid. This means highly content-valid. C-OAR-SE theory reveals that most if not all of our measures of soft constructs have unacceptably low content validity. The problem applies especially to measures of abstract (multiple-item) constructs but also to unnecessary multiple-item measures of concrete (single-item) constructs.

The discomfiting conclusion to be drawn from this article is that most of our measured knowledge in the social sciences, including marketing, is questionable, and that a not unsubstantial amount of this knowledge – especially the recent

"Churchill-based" knowledge – is wrong. There is no doubt that our whole approach to measurement needs rethinking.

References

Armstrong, J.S. (1967), "Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine", *The American Statistician*, Vol. 21 No. 5, pp. 17-21.

Armstrong, J.S. and Soelberg, P. (1968), "On the interpretation of factor analysis", *Psychological Bulletin*, Vol. 70 No. 5, pp. 361-4.

Bagozzi, R.P. (1994), "Structural equation models in marketing research: basic principles", in Bagozzi, R.P. (Ed.), *Principles of Marketing Research*, Blackwell, Cambridge, MA, pp. 317-85.

Bearden, W.O. and Netemeyer, R.G. (1999), *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, Sage, Thousand Oaks, CA.

Bergkvist, L. and Rossiter, J. (2007), "The predictive validity of multiple-item vs single-item measures of the same constructs", *Journal of Marketing Research*, Vol. 44 No. 2, pp. 175-84.

Borsboom, D. (2005), *Measuring the Mind*, Cambridge University Press, Cambridge.

Brakus, J.J., Schmitt, B.H. and Zarantonello, L. (2009), "Brand experience: what is it? How is it measured? Does it affect loyalty?", *Journal of Marketing*, Vol. 73 No. 2, pp. 52-68.

Campbell, D.T. and Fiske, D.W. (1959), "Convergent and discriminant validation by the multitrait-multimethod matrix", *Psychological Bulletin*, Vol. 56 No. 2, pp. 81-105.

Churchill, G.A. Jr (1979), "A paradigm for developing better measures of marketing constructs", *Journal of Marketing Research*, Vol. 16 No. 1, pp. 64-73.

Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences*, revised ed., Academic Press, New York, NY.

Cronbach, L.J. (1951), "Coefficient alpha and the internal structure of tests", *Psychometrika*, Vol. 16 No. 3, pp. 297-334.

Cronbach, L.J. (1961), *Essentials of Psychological Testing*, 2nd ed., Harper & Row, New York, NY.

Deshpandé, R. and Webster, F.E. (1989), "Organizational culture and marketing: defining the research agenda", *Journal of Marketing*, Vol. 53 No. 1, pp. 3-15.

Deshpandé, R., Farley, J.U. and Webster, F.E. (1993), "Corporate culture, customer orientation, and innovation in Japanese firms: a quadrant analysis", *Journal of Marketing*, Vol. 57 No. 1, pp. 23-37.

Diamantopoulos, A. and Sigauw, J.A. (2006), "Formative versus reflection indicators in organizational measure development: a comparison and empirical demonstration", *British Journal of Management*, Vol. 17 No. 4, pp. 263-82.

Ehrenberg, A.S.C. (1975), *Data Reduction: Analysing and Interpreting Statistical Data*, Wiley, London.

Gelfand, M.J., Erez, M. and Ayean, Z. (2007), "Cross-culture organizational behavior", *Annual Review of Psychology*, Vol. 58, pp. 479-534.

Hardesty, D.M. and Bearden, W.O. (2004), "The use of expert judges in scale development", *Journal of Business Research*, Vol. 57 No. 2, pp. 98-107.

Homburg, C., Wieseke, J. and Bornemann, T. (2009), "Implementing the marketing concept at the employee-customer interface: the role of customer need knowledge", *Journal of Marketing*, Vol. 73 No. 4, pp. 64-81.

Kelley, T.L. (1927), *Interpretation of Educational Measurements*, Macmillan, New York, NY.

Krauss, S.J. (1995), "Attitudes and the prediction of behavior: a meta-analysis of the empirical literature", *Personality and Social Psychology Bulletin*, Vol. 21 No. 1, pp. 59-75.

Leeflang, P.S.H., Bijmolt, T.H.A., van Doorn, J., Hanssens, D.M., van Heerde, H.J., Verhoef, P.C. and Wierenga, J.E. (2009), "Creating lift versus building the base: current trends in marketing dynamics", *International Journal of Research in Marketing*, Vol. 26 No. 1, pp. 13-20.

Lord, F.M. and Novick, M. (1968), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA.

McGuire, W.J. (1989), "The structure of individual attitudes and attitude systems", in Pratkanis, A.R., Breckler, S.J. and Greenwald, A.G. (Eds), *Attitude Structure and Function*, Erlbaum, Mahwah, NJ, pp. 37-68.

Mosteller, F. and Youtz, C. (1990), "Quantifying probabilistic expressions", *Statistical Science*, Vol. 5 No. 1, pp. 2-34.

Narver, J.C. and Slater, S.F. (1990), "The effect of market orientation on business profitability", *Journal of Marketing*, Vol. 54 No. 4, pp. 20-35.

Nunnally, J.C. (1978), *Psychometric Theory*, 2nd ed., McGraw-Hill, New York, NY.

Osgood, C.E., Suci, G.J. and Tannenbaum, P.H. (1957), *The Measurement of Meaning*, University of Illinois Press, Urbana, IL.

Ouellette, J.A. and Wood, W. (1998), "Habit and intention in everyday life: the multiple processes by which past behavior predicts future behavior", *Psychological Bulletin*, Vol. 124 No. 1, pp. 54-74.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1994), "Alternative scales for measuring service quality: a comparative assessment based on psychometric and diagnostic criteria", *Journal of Retailing*, Vol. 70 No. 3, pp. 201-30.

Peterson, R.A., Albaum, G. and Beltramini, R.F. (1985), "A meta-analysis of effect sizes in consumer behavior experiments", *Journal of Consumer Research*, Vol. 12 No. 1, pp. 97-103.

Revelle, W. (1979), "Hierarchical clustering and the internal structure of tests", *Multivariate Behavioral Research*, Vol. 14 No. 1, pp. 19-28.

Rossiter, J.R. (2001), "What is marketing knowledge? Stage 1: forms of marketing knowledge", *Marketing Theory*, Vol. 1 No. 1, pp. 9-26.

Rossiter, J.R. (2002a), "The C-OAR-SE procedure for scale development in marketing", *International Journal of Research in Marketing*, Vol. 19 No. 4, pp. 305-35.

Rossiter, J.R. (2002b), "The five forms of transmissible, usable marketing knowledge", *Marketing Theory*, Vol. 2 No. 4, pp. 369-80.

Rossiter, J.R. (2005), "Reminder: a horse is a horse", *International Journal of Research in Marketing*, Vol. 22 No. 1, pp. 23-5.

Rossiter, J.R. (2007), "Toward a valid measure of e-retailing service quality", *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 2 No. 3, pp. 36-48.

Rossiter, J.R. (2008), "Content validity of measures of abstract constructs in management and organizational research", *British Journal of Management*, Vol. 19 No. 4, pp. 380-8.

Rossiter, J.R. (2009a), "ER-SERVCOMPSQUAL: a measure of e-retailing service components quality", *Service Science*, Vol. 1 No. 4, pp. 212-24.

Rossiter, J.R. (2009b), "Qualitative marketing research: theory and practice", *Australasian Journal of Market and Social Research*, Vol. 17 No. 1, pp. 7-27.

Rossiter, J.R. (2011), *Measurement for the Social Sciences: The C-OAR-SE Method and Why it Must Replace Psychometrics*, Springer, New York, NY.

Rossiter, J.R. and Bellman, S. (2005), *Marketing Communications: Theory and Applications*, Pearson Prentice Hall, Sydney.

Rossiter, J.R. and Bergkvist, L. (2009), "The importance of choosing one good item for single-item measures and its generalization to all measures", *Transfer: Werbeforschung & Praxis*, Vol. 55 No. 2, pp. 8-18.

Rossiter, J.R. and Percy, L. (1987), *Advertising and Promotion Management*, McGraw-Hill, New York, NY.

Rossiter, J.R. and Percy, L. (1997), *Advertising Communications & Promotion Management*, McGraw-Hill, New York, NY.

Spearman, C. (1904), "General intelligence, objectively determined and measured", *American Journal of Psychology*, Vol. 15 No. 2, pp. 201-93.

Tellis, G.J., Prabhu, J.C. and Chandy, R.K. (2009), "Radical innovation across nations: the preeminence of corporate culture", *Journal of Marketing*, Vol. 73 No. 1, pp. 3-23.

Urban, G.L. and Star, S.H. (1991), *Advanced Marketing Strategy*, Prentice Hall, Englewood Cliffs, NJ.

Viswanathan, M., Sudman, S. and Johnson, M. (2004), "Maximum versus meaningful discrimination in scale response: implications for validity of measurement of consumer perceptions about products", *Journal of Business Research*, Vol. 47 No. 1, pp. 108-25.

Corresponding author
John R. Rossiter can be contacted at: john_rossiter@uow.edu.au