The evolution of “classical mythology” within marketing measure development

Nick Lee and Graham Hooley
Aston Business School, Aston University, Birmingham, UK

Abstract

Purpose – This paper provides a deeper examination of the fundamentals of commonly-used techniques – such as coefficient alpha and factor analysis – in order to more strongly link the techniques used by marketing and social researchers to their underlying psychometric and statistical rationale.

Design/methodology approach – A wide-ranging review and synthesis of psychometric and other measurement literature both within and outside the marketing field is used to illuminate and reconsider a number of misconceptions which seem to have evolved in marketing research.

Findings – The research finds that marketing scholars have generally concentrated on reporting what are essentially arbitrary figures such as coefficient alpha, without fully understanding what these figures imply. It is argued that, if the link between theory and technique is not clearly understood, use of psychometric measure development tools actually runs the risk of detracting from the validity of the measures rather than enhancing it.

Research limitations/implications – The focus on one stage of a particular form of measure development could be seen as rather specialised. The paper also runs the risk of increasing the amount of dogma surrounding measurement, which runs contrary to the spirit of this paper.

Practical implications – This paper shows that researchers may need to spend more time interpreting measurement results. Rather than simply referring to precedence, one needs to understand the link between measurement theory and actual technique.

Originality/value – This paper presents psychometric measurement and item analysis theory in easily understandable format, and offers an important set of conceptual tools for researchers in many fields.

Keywords Market research, Psychometric tests, Statistics

Paper type Research paper

Introduction

Since 1979, when Churchill’s seminal article “A paradigm for measure development in marketing research” (Churchill, 1979) was published in the Journal of Marketing Research, the use of complex statistical tools in the development of reflective, multi-item scales to measure latent constructs within marketing research has flourished. In the decade between 1980 and 1989 Bruner and Hensel (1993) reported 750 instances of multi-item scales in the six leading marketing journals. It cannot be denied that the widespread adoption of this “classical” measure development paradigm has enriched marketing science immeasurably as a discipline, over and above its rather unrefined state prior to 1979 (see Peter, 1979). More recently, the level of sophistication of marketing measure development has been further enhanced by the introduction into non-technical literature of confirmatory factor analysis (CFA) methodology by authors such as Gerbing and Anderson (1988), Gerbing and Hamilton (1996), Bagozzi and
Yi (1988), and Fornell and Larcker (1981). The great benefit to a field such as marketing of articles such as those referred to above, is that they distill critical methodological knowledge in such a way as to enable its widespread adoption by those not inclined to delve into complex, obscure, and often impenetrable technical resources. In such ways is the evolution of marketing as a science assisted since, as an applied discipline, marketing scholars (if most of us are honest) are primarily interested in the results of studies, not the intricacies of analysis and mathematics involved in obtaining those results.

However, the enormous benefits of following paradigms such as that proposed by Churchill (1979) have also masked some of their dangers to marketing scholarship, which are inherent to the latter’s applied nature. In particular, the great efforts expended by authors such as Churchill (1979), Gerbing and Anderson (1988), Bagozzi and Yi (1988), and Diamantopoulos (1999) to make difficult and rigorous methods accessible to general marketing scholarship, seem to have resulted in the development of a number of misconceptions, or myths, around the application of these complex methods. It could be argued that these misconceptions have evolved as a result of our increasing ability to use complex methods, without the requirement of a fundamental understanding of their mathematical and – crucially – conceptual underpinnings.

Furthermore, these myths and misconceptions tend to be perpetuated within marketing research since it is a field where one of the primary justifications for any method’s use is simple precedent (Stewart, 1981). To put things more succinctly, the labours of the distinguished authors already quoted (and many others) have allowed marketing academics to avoid dealing with a number of critical considerations simply by quoting or referring back to their work. Furthermore, the continued lack of attention to these issues has led to a state where analysis techniques are applied slavishly to a given task to which they may be inappropriate, or the results may lack sufficient explanation to assure readers of their appropriateness. As a consequence there may be serious problems with the measures we develop, and ultimately the theories we aim to test. In fact, this has been stated even more strongly by one of the “fathers” of psychometric testing, Paul Kline: “to follow the precepts of [test construction and/or use] without understanding their psychometric rationale is to accept dogma, the antithesis of a scientific approach to . . . testing, a cast of mind which in earlier times led inevitably to the dark ages” (Kline, 2000, p. 32).

This paper is an attempt to go some way towards addressing the above concerns within the context of the evolution of marketing research. More specifically, the aim of this paper is to present some fundamental issues concerning classical measure development techniques in a non-technical manner and to a wide audience of general marketing academics. The issues we detail are concerned solely with the classical psychometric paradigm for measure development as defined by authors such as Churchill (1979), DeVellis (1991) and Spector (1992), and the techniques we discuss are based around scale purification, reliability and validity assessment. While we recognise that contemporary work has expanded the canon of techniques available to marketing research (e.g. Diamantopoulos and Winkhofer, 2001; Gerbing and Anderson, 1988; Rossiter, 2002), we also recognise that a great many of our existing measures were based upon Churchill’s (1979) construction method, and indeed many still continue to be. Furthermore, by discussing methods that have had a long history within marketing, we hope to illuminate critical evolutionary issues concerning our use
of classical measure development methods, which may not be as prevalent when discussing newer methods without such a history of usage. Nevertheless, it should be clear that our aim is not to present a technical explication but a conceptual one, and to utilise well-founded marketing methodologies to illustrate how the evolution of marketing research has proceeded to a state whereby it is arguable that our use of classical measure development methods may in fact have detracted from our goal of providing valid, reliable measures.

We begin with a brief reiteration of the “classical” paradigm for measure development introduced into marketing by Churchill (1979), with particular focus on the purification stage of the process. Following this, the rest of the paper discusses issues related to two key methods of measure purification. First, we discuss some key potential problems with the concept of internal consistency in purifying measures, and of Cronbach’s (1951) coefficient alpha to measure internal consistency. We also suggest some areas where we could improve our use of internal consistency and alpha criteria. Following this, we then turn to issues concerned with the usage of exploratory factor analysis (EFA) in measure purification, and suggest areas for particular caution for marketing scholars. Finally, we provide some conclusions and suggest some critical ways in which marketing researchers can advance the use of existing measure development tools in achieving the goals of scientific measure development.

The “classical” paradigm for the development of measures of marketing constructs

The publication of Churchill’s seminal paper on construct development marks a watershed in the evolution of marketing science. Prior to the 1970s many marketing measures were poorly developed and poorly tested (Jacoby, 1978; Peter, 1979), and Churchill’s paper was essentially the first to apply classical psychometric theory (which was then already well-accepted in other disciplines) to the marketing discipline. In the introduction to his paper, Churchill (1979) quoted a then-recent paper by Jacoby (1978, p. 91):

More stupefying than the sheer number of our measures is the ease with which they are proposed and the uncritical manner in which they are accepted. In point of fact, most of our measures are only measures because someone says they are, not because they have been shown to satisfy standard measurement criteria (validity, reliability and sensitivity).

In proposing a more systematic and rigorous approach to measure development, Churchill laid the groundwork and procedures for the next generation of marketing scholars. Specifically, Churchill argued for the development of multi-item measures as more robust ways to measure marketing phenomena, and put forward an eight-step procedure to achieve that. This classical paradigm is presented in Figure 1. While few would argue with the stages proposed, the issue of how each stage is undertaken, and what techniques are appropriate for each, is open to question (summarised in the right-hand column in Figure 1). This is because the stages are conceptual, yet we must use various specific operational techniques to achieve their goals. If the link between the operationalisation and the conceptual goal is broken, we run the risk of hampering rather than helping our tasks. In particular, the central stage of “measure purification” is an area where technical dogma and tradition have dominated marketing academia, to the potential detriment of better measures.
Measure purification

Classical psychometrics (see Nunnally, 1967), as advocated by Churchill (1979), recommends that marketing measures be developed through the systematic generation and subsequent “purification” of a large bank of scale items (DeVellis, 1991; Spector, 1992). Under this approach, the underlying construct that the researcher seeks to measure is represented by a “sample” of the hypothetical, and usually infinite, set of all the items which could possibly be created to measure the construct (see Kline, 2000). The critical issues in identifying an appropriate sample of items are:

- to ensure that the final set of items used captures or reflects the underlying construct as fully as possible, but without redundancy; and
- to ensure that the items really do reflect the underlying construct and not some other, similar but distinct, construct.

The scale-purification process attempts to “purify” the scale (after the initial item-generation and data-collection phases) to remove redundant or non-reflective items. Two statistical techniques have traditionally been recommended for scale purification, coefficient alpha and EFA (Churchill, 1979). The balance of the present

Figure 1. Churchill’s paradigm for developing better measures of marketing constructs

Source: Churchill (1979)
paper discusses a number of concerns with both techniques, which the evolution of marketing research appears to have marginalised, to the detriment of the conceptual goals of the classical measure development paradigm. A glossary in the Appendix at the end of this paper provides definitions of terms used.

Cronbach’s coefficient alpha and the reliability of multi-item scales
When developing multi-item scales according to the classical paradigm, Cronbach’s (1951) coefficient alpha had been the measure of choice within marketing and many other fields for assessing the reliability of a given scale (Peter, 1979; Peterson, 1994). While the “composite reliability” coefficient (Fornell and Larcker, 1981; Gerbing and Anderson, 1988) has received some use recently, coefficient alpha still appears to be the primary foundation upon which the reliability assessment of marketing scales is based. However, while coefficient alpha has close to an unassailable position within marketing research, scholars in other disciplines have raised a number of concerns over the use and misuse of alpha, even prior to Churchill’s (1979) work (e.g. Boyle, 1991; Cattell, 1973; Cattell and Kline, 1977; Cortina, 1993; Green et al., 1977; Hattie, 1985; Rogers et al., 2002). Indeed, even within marketing it has been suggested that there is some need for care when using coefficient alpha to assess reliability, although no real criticisms appear to have been levelled until recently (e.g. Rossiter, 2002). However, before discussing the possible limitations of coefficient alpha as a reliability indicator, and any subsequent caveats for its usage, a brief re-examination of both the concept of measure reliability, and the coefficient alpha statistic itself is presented.

Reliability and validity
The reliability of a measure is high when “independent but comparable measures of the same trait or construct of a given object agree” (Churchill, 1979, p. 65). Reliability is unimportant in and of itself, but is nevertheless vital since in order to be valid (i.e. to measure what it claims to measure) a given measure must be reliable (Kline, 2000; Peter, 1979; Spector, 1992). It is important to note, however, that a measure may be reliable without being valid (Kline, 2000).

Reliability is concerned with the proportion of variance in a measure, which is attributable to the true score on the latent construct that is being measured (DeVellis, 1991). For example, if variance in a measure of say, job satisfaction, is comprised of a high proportion attributable to the true score on the latent job satisfaction construct, then that measure of job satisfaction is reliable. Conversely, if a subject’s score on the job satisfaction measure is comprised of a high percentage of random error, then the job satisfaction measure is not reliable. Thus, reliability is essentially a theoretical concept, in that (practically speaking) we can never be certain of the true score of a subject on a given latent construct (Kerlinger and Lee, 2000) – and if we could then we would measure the latent construct directly, and it would no longer be latent. Furthermore, we can never be certain that the other variance in a measure not attributable to the latent construct is in fact all random error and not some part of systematic error (such as respondent biases). As a consequence of this definition, it is clear that reliability can never be exactly determined, but it can be approximated in a number of ways (see Kerlinger and Lee, 2000; Kline, 2000).

Drawing from classical measurement theory, it has been suggested that if the scores of the items of a reflective scale are highly influenced by the level of the latent construct
they are designed to measure (in other words not influenced by error), then the items
should also have strong relationships (i.e. high intercorrelations) with each other
(DeVellis, 1991). Thus, high intercorrelations between scale items should indicate that
the items are also strongly related to the latent construct, and thus not highly
influenced by random error. Scales that have high intercorrelations between the items
are known as internally consistent (see Kline, 2000; DeVellis, 1991; Churchill, 1979).

Internal consistency
The internal consistency of a scale is generally measured by coefficient alpha, although
this is by no means the only workable approach (Kline, 2000; Spector, 1992). Coefficient
alpha is, practically speaking, “the proportion of a scale’s total variance that is
attributable to a common source” (DeVellis, 1991, p. 27), nothing more, nothing less. By
definition, this means that alpha is also a measure of the unique variance in a scale.
However, it should be clear from the above discussion that coefficient alpha is most
definitely not an actual measure of reliability. Furthermore, while the “common source”
referred to by DeVellis (1991) is assumed to be the latent construct we are trying to
measure, this is only theoretically the case, and the link between the latent construct
and the actual scale items is therefore purely theoretical. However, coefficient alpha can
be utilised in estimating the reliability of a multi-item reflective scale, by providing an
indication of a scale’s internal consistency. Nevertheless, the characteristics of both
internal consistency, and coefficient alpha, necessitate a number of caveats.

While most psychometricians are in no doubt that internal consistency is of
considerable utility in estimating the reliability of a measure (see Kline, 2000), a
number of dissenters do not share this confidence. In particular, concern has been
heard for a significant amount of time from Cattell (e.g. Cattell, 1973; Cattell and Kline,
1977), while more recent caution has come from Boyle (1991) and Rossiter (2002) among
others. The root source of this concern appears to be the possible negative impact of
high levels of internal consistency on the validity of a measure. Drawing from the
earlier discussion, if reliability is primarily important because of its impact on measure
validity, then for internal consistency to be a good indicator of reliability, logically
speaking it must not negatively impact a measure's validity.

There are two key premises to the argument against internal consistency. First,
high internal consistency is strongly related to high intercorrelations between items.
Second, for a scale to be a valid and useful measure of a latent construct, the scale
scores must be able to predict the latent true scores to a high degree (Kline, 2000). Of
course, according to classical reflective scale theory, technically speaking it is the latent
true score that should predict the scale item scores. However, since we can never
actually measure the true score of a latent construct, this distinction is arguably purely
academic. Furthermore, if the true score is in theory able to predict the scale scores
accurately, then in practical terms the reverse should also be true. Notwithstanding the
above technicalities, drawing from multicollinearity theory (which for marketing
researchers has previously been primarily concerned with multiple regression), it can
be shown that if we wish to predict the value of any variable from a set of independent
items, then the items should not be correlated with each other (see Kleinbaum et al.,
1998). As the correlation between the items increases, the explanatory power of any
single item decreases (e.g. Kleinbaum et al., 1998; Hair et al., 1998). While this point has
received some discussion in literature relating to construction of formative indices
within marketing (e.g. Diamantopoulos and Winklhofer, 2001), marketing-focused theory pertaining to classical reflective scales has not dealt with the issue.

Nevertheless, Kline (2000, p. 12) suggests that “a test [i.e. a reflective scale] can be seen as a set of items with which we intend to predict the criterion [i.e. the latent construct] test score”. In other words for a scale to be valid, a subject’s score on a multi-item scale should be closely predictive of the subject’s actual level of the unobservable construct the scale is designed to measure. Thus, it would follow that the scale would be best able to predict this latent construct score if each individual item was highly correlated with the true latent construct score but correlated zero with all the other items. While such a scale would have a high level of validity (since it is able to measure what it claims to measure with a high accuracy), it would have a very low level of internal consistency, and thus a low coefficient alpha. By contrast, a scale with a high level of internal consistency is likely to have a high level of intercorrelation between its items, which would suggest that a number of those items may in fact be redundant in predicting the subject’s true score on the latent construct (see Boyle, 1991; Cattell, 1973). In other words, scales high in internal consistency may be comprised of items that are essentially paraphrases or repetitions of each other (Kline, 2000; Smith, 1999). Put crudely, a scale consisting of the same item repeated six times will give an excellent alpha (of 1.0 if respondents are consistent in their answers) but will have low validity in measuring the latent construct.

This can lead to a situation where a scale has significant item redundancy, and may actually tap only a narrow area of a latent construct, leaving it low in validity (Boyle, 1991). Therefore, researchers should clearly be cautious when using internal consistency as a proxy for reliability in their scale development activities. Dogmatic adherence to high internal consistency can lead to high levels of item redundancy and measures which tap only limited aspects of a latent construct, ultimately leaving them of low validity.

Notwithstanding the above issues, internal consistency remains a valuable source of information about a scale, whether that information suggests high reliability or a worrying level of item redundancy. It is the researcher’s job to interpret that information. However, if one wishes to evaluate internal consistency in scale development, then the characteristics of coefficient alpha must be understood also, in order for scale developers and users to interpret the results of a coefficient alpha test.

**The effects of scale length on coefficient alpha**

Coefficient alpha also has some relatively well-known caveats on its usage, perhaps the best known of which concerns the number of items in the scale to be tested. Specifically, as an artefact of the alpha formula itself, coefficient alpha will increase as the number of scale items increases ceteris paribus (Hair *et al*., 1998). So in other words, regardless of the level of internal consistency achieved by a scale, if enough items are present the alpha will be high. In theory this is not a problem, since it should be expected that as the number of items increases, so should the reliability of the scale as it gets closer to including the theoretical “universe” of all possible items (Kline, 2000). Despite the latter, in practical terms a problem may exist since the universe of all possible items is usually infinite (Kline, 2000). Thus, researchers must be wary if measures of reliability such as alpha are influenced by scale lengths, which in no way approach this universe.
Peterson’s (1994) meta-analysis discovered that scale length had a significant impact on coefficient alpha ($R^2$ of 0.1), a result also achieved by Churchill and Peter’s (1984) marketing-based research. While in absolute terms Peterson (1994, p. 386) suggested that this impact was not “especially strong”, it should be considered that a systematic bias of around 10 per cent (i.e. an $R^2$ of 0.1) is likely to have a substantive impact on the assumed reliability and validity of a measure. Unfortunately, further research outside of marketing has found that scale length has an even more pervasive effect on coefficient alpha. For example Cortina (1993) found that he could achieve levels of alpha commonly regarded as sufficient (i.e. over 0.7) with scales of 12 items and above even when the scale consisted of two or more completely uncorrelated dimensions. Thus, it can be seen that coefficient alpha can be problematic as a measure of internal consistency when scales consist of practically large numbers of items (i.e. well before such numbers approach the theoretical levels where reliability should significantly increase). In such situations, large values of alpha do not necessarily indicate either the presence of a single, homogenous factor (i.e. internal consistency), or high reliability, as is commonly assumed (see DeVellis, 1991).

Furthermore, as might be expected from the previous discussion, coefficient alpha is downwardly biased when few items are present in a scale. In fact, the relationship between item numbers and alpha score is curvilinear (Komorita and Graham, 1965), which perhaps led to Peterson’s (1994) suggestion that the bias at low item numbers is more pronounced than at high. More specifically, it has been previously argued that, with very small numbers of scale items, coefficient alpha scores will underestimate internal consistency of the scale (see Gerbing and Anderson, 1988). Within marketing research this is a significant issue. Specifically, respondents to large-scale marketing surveys will not in general tolerate large questionnaires, and thus marketing researchers should try to reduce the length of their instruments wherever possible (see Churchill, 1999). Furthermore, increased use of covariance structure analysis methods seems likely to increase reliance on short scale lengths in order to reduce the number of estimated parameters (see Diamantopoulos and Siguaw, 2000; Kelloway, 1998; Sharma, 1996). As a result of this, the use of coefficient alpha under conditions of sub-optimal scale length should be considered very carefully. However, the use of the composite reliability formula (e.g. Gerbing and Anderson, 1988) can avoid problems regarding scale length, among other issues (Fornell and Larcker, 1981).

Multidimensionality of scales
A further issue concerning coefficient alpha, which was implied above, is its performance under conditions of multidimensionality. From the theories discussed so far, it is clear that coefficient alpha should be a measure of the extent to which a hypothesised common factor is present in all items. Furthermore, if there is one common factor to a scale, alpha should be a measure of the strength of that factor (Cortina, 1993). Therefore, if a scale is comprised of more than a single dimension, then to be a useful measure of internal consistency, alpha should provide some indication of this fact. Unfortunately, a significant amount of research has provided evidence that this is not the case (e.g. Cortina, 1993; Gerbing and Anderson, 1988; Green et al., 1977). In fact if enough items are present, a scale can contain even five separate dimensions of which each item loads on two separate dimensions and still return an alpha greater
than 0.8 (Green et al., 1977). The upshot of this is that high coefficient alphas cannot be taken as indicators of unidimensionality in a given scale (e.g. Gerbing and Anderson, 1988). This is because as mentioned earlier, high coefficient alphas simply measure the amount of unique and common item variance in a scale, and not what that common variance consists of. As alluded to earlier, the make-up of the common variance in a given scale (i.e. exactly what construct that scale measures) is a conceptual decision made by the researcher, not something that can be drawn from the results of a mathematical operation.

The fact that high alphas do not indicate unidimensionality is unfortunate in the context of marketing research. In particular, Churchill (1979, p. 68, emphasis in original) explicitly stated that “[c]oefficient alpha absolutely should be the first measure one calculates to assess the quality of the [scale]”. This statement seems to imply that, if coefficient alpha reaches an acceptable level, one should simply take a measure to be of high quality. However, the theory presented above suggests clearly that this should not be the case. Nevertheless, despite the introduction of a number of sophisticated measure development techniques which do provide excellent indications of measure dimensionality (e.g. Gerbing and Anderson, 1988, etc.), it has been quite acceptable to utilise Churchill’s (1979) process even in the recent past. Furthermore, whether or not this changes in the future, it is undeniable that many of our key measures have been developed using the latter methodology, and are often repeatedly used by marketing scholars in a relatively unquestioning manner.

Summary
If one were to reflect on the arguments presented so far, one could rather easily get the impression from the above that coefficient alpha should be avoided at all costs. However, this is not our recommendation. Instead, we strongly advocate the interpretation of coefficient alpha values for scales, rather than the unquestioning reporting of scale alphas, usually coupled with the rigid acceptance of anything over 0.7. It is apparent that an alpha of 0.7 may technically mean exactly the same thing across different scales (i.e. the proportion of unique item variance), but in practical terms mean wildly different things, and in some cases mean exactly nothing of practical use, such as when item numbers are very large. Thus, marketing researchers should, and should have in the past, taken into account a number of factors when interpreting coefficient alpha:

1. Most important, researchers need to understand exactly what they are looking at, in terms of both internal consistency and alpha itself. In other words, does one actually want to measure internal consistency, and should it be high or low, and what do high/low alphas actually mean in the context of the study?

2. If alpha is used, researchers should be aware of the specific factors which can influence alpha, such as item numbers.

3. Researchers need to realise that if dimensionality is an issue (and according to classical theory it always is), one cannot rely solely on an alpha value, but must at minimum compute a precision estimate (see Cortina, 1993), and ideally utilise FA to examine dimensionality more fully.

4. Finally, researchers should interpret alpha scores in light of the factors above, and also in light of alphas which have been reported for the scale when it has
been used in other studies. For example, an alpha of 0.7 on a job satisfaction scale may require different levels of concern depending on alphas reported in previous uses of the scale.

The use of EFA in scale development
While internal consistency and coefficient alpha are often primary tools in the early development of measures, EFA also assumes a prominent place in the measure purification process (Churchill, 1979; Gerbing and Anderson, 1988). However, it has been argued that there are many misconceptions about FA which have developed within marketing research, and that these myths have become so integrated into marketing research’s imaginary that such misinformation is perpetuated even by “[factor analysis] defenders and some prominent reviewers” (Stewart, 1981, p. 51). Particular misapplications of EFA are many and varied, and indeed some marketing-focused literature has made attempts to provide discussion of a number of misconceptions (e.g. Flynn and Pearcy, 2001; Stewart, 1981). However, there remain important issues in need of substantive explication and justification. Some of these issues are relevant to marketing research as it has evolved, and others are especially pertinent due to newer developments in marketing science such as the introduction of alternative theories of measurement (e.g. Diamantopoulos and Winklhofer, 2001; Rossiter, 2002).

Principal components analysis (PCA) and FA
The first issue is the potential confusion which may arise between PCA and FA. Scholars outside of marketing have noted that there is considerable misunderstanding over the differences between PCA and FA, although they are conceptually and technically quite distinct (e.g. Sharma, 1996). However, the critical issues regarding these differences do not appear to have had significant discussion within marketing research literature. This has led to a situation where marketing researchers may employ PCA to do the (quite different) job of FA for a number of reasons. First, PCA and FA are both techniques of data reduction in that, ostensibly at least, they both reduce large numbers of items down to a smaller number of components or factors. Second (perhaps due to this similarity), many popular statistical software packages allow one to perform both PCA and FA from the same sub-menu, and indeed in some, PCA is the default option. The confusion is further exacerbated by the existence of principal components factoring, which is essentially a special case of PCA and arguably not actually FA at all (Sharma, 1996).

Dealing with the conceptual differences first, it should be clearly stated at the outset that the aims of PCA and FA are quite different. The objective of PCA is to utilise the observed variance in the data set to create new variables which are composed of the original items. The objective of FA is to identify an underlying or latent factor which is responsible for observed correlations among the original items (see Kline, 2000; Sharma, 1996). Drawing from this, it can be seen that components extracted from a PCA do not mean anything conceptually, other than what they are composed of. In other words, they are an index of the original items, and a given component is simply a linear combination of the relevant items (see Bollen and Lennox, 1991). By contrast, a factor extracted from a FA is by definition responsible for the correlation between the relevant items, and thus does represent an underlying common or latent factor. Therefore, if one is to return to classical scale development theory, the use of PCA to
determine the underlying factor structure (i.e. the dimensionality) of a multi-item reflective scale is conceptually inappropriate, since the scale items in fact form a principal component, rather than reflecting the presence of an underlying factor (as is the case with a FA). Regardless of the conceptual differences between PCA and FA, it has been suggested that their actual results may be relatively similar within a given data set, particularly with large numbers of cases (see Stewart, 1981; Sharma, 1996), and thus it has been implied that in practical terms researchers should not be too concerned which technique they use (Stewart, 1981). However, without going into technical detail it can be seen that the variables that together make up a component formed by PCA are not assumed to have any unique variance (i.e. random error). This is due to the fact that the component is a straightforward linear combination of the items. However, within FA, each item is assumed to consist of a segment of shared variance, and a segment of unique variance (Sharma, 1996), with the shared variance representing the presence of the latent common factor. Moving back to the principles of classical measure development (see Churchill, 1979), it can be seen that an assumption is made that any single scale item’s observed score is comprised of a component attributable to the latent construct, and also of a unique component (perhaps caused by measurement error, or other factors). Therefore, utilising PCA to determine the underlying dimensionality of a reflective scale is fundamentally misleading and incorrect. In fact, in purely statistical terms, since error or other unique variance is not factored in, we should expect the loadings of individual items on the principal components in PCA to be inflated when compared with their counterparts in FA (Stewart, 1981), which could lead to errors and inconsistencies in scale construction when PCA is used for the latter purpose. Furthermore since, by using PCA to develop reflective scales, we are not taking into account the measurement or other error which classical theory tells us always exists in a given set of items, there is no reason to expect the “dimensions” uncovered by a PCA to be generalisable to other samples. This is because the components formed by a PCA are completely specified by the values of the individual items, and can be argued to be simply artefacts of the data set used to create them (see Kline, 2000; Nunnally, 1967).

**Ejection of items due to low communality**

Communality is the amount of variance an original variable shares with all other variables included in a FA (Hair et al., 1998). A communality of under 0.5 signifies that less than half of the variance in the item has been taken into account in identifying the latent construct (Hair et al., 1998). In psychometric terms, we consider the remaining variance to be made up of specific (also known as unique) variance and error variance. Specific variance is that variance associated with another variable, while error variance is that variance due to unreliability in the data-gathering process, measurement error or a random component in the measured phenomenon (Hair et al., 1998). It is common practice in scale purification through FA automatically to eject items from the analysis which exhibit low communality (e.g. Savery, 1993; Singh, 2000).

However, drawing from the earlier discussion of internal consistency, it can be seen that the practice of ejecting items exhibiting low communality should be done with care. For example, a highly internally consistent measure containing a set of essentially redundant items will likely have high communalities for the items. By contrast, the
more valid yet less internally consistent measures advocated by authors such as Cattell (1973) are likely to have significantly lower item communalities. As a result, the rejection of items with low communalities can only be conceptually justified where it is either suspected that the error variance is high, or that the item may be tapping a substantively different construct or dimension. However, if the researcher suspects that this is not the case, a situation where common variance is low, but specific, or unique variance is high this may mean that the item is the sole measure of a particular aspect or nuance of a construct. To eject such an item could result in the loss of an important, and unique, contributing item. Indeed, presence of low communality in an item generated through sound prior research (literature based or empirical) should signal the possibility of under-specification of the particular aspect of the overall conceptual domain tapped by the item.

Unfortunately there are no simple, statistical means for differentiating unique variance from error variance. It is therefore incumbent on the analyst to go back through the previous stages of the research (i.e. the item generation process in the Churchill model) to ensure that the item was well specified. Only where there can be confidence that either the error variance is high, or the item does not tap the appropriate conceptual domain, should the item be ejected from further analysis.

Selection of number of factors
A related issue concerns the number of factors to extract in the FA. In EFAs in marketing the common practice is to use the “Kaiser criterion” to make this decision (e.g. Alexander and Colgate, 2000; Urban, 2001). Under this approach factors are extracted only if they explain more of the total variance than a single item is capable of explaining. Each variable contributes a value of 1 to the total eigenvalue, hence any factors with an eigenvalue of less than 1 are discounted (Kim and Mueller, 1978). However, it is interesting to note that – technically speaking – the Kaiser criterion is only applicable when using PCA, although it has been argued that it is “applicable in spirit” to other types of FA (DeVellis, 1991, p. 97). In conceptual terms this is because the Kaiser criterion is solely based on maximising the variance explained of each factor (the goal of PCA), rather than their conceptual value (more in tune with the goal of FA). Thus researchers using FA should naturally be cautious when applying the Kaiser criterion in all situations. However in more specific terms, as Hair et al. (1998) point out, the use of the Kaiser criterion as a cut-off is most reliable when the number of variables is between 20 and 50. With less than 20 variables there is a tendency for this method to extract a conservative number of factors, whereas if the number of variables is greater than 50 too many factors may be extracted.

Furthermore, it can be suggested that items which exhibit low levels of communality are also unlikely to appear in factors with eigenvalues greater than 1, since they exhibit low levels of correlations with all of the other items. In fact, what is likely to happen is that such items will again be candidates for ejection. Again, it is incumbent on the analyst to interpret this result carefully by returning to the earlier specification stages of the research to ensure that the items are well specified.

Taken together, the issues raised concerning items with low communalities, or which load only on factors with eigenvalues less than 1, suggest the need to re-visit the specification stages of the research, rather than attempt to “correct” problems in the purification stage (one could perhaps refer to the old adage “garbage in – garbage out”
at this point). These might be poorly specified items with high error variance that deserve to be ejected if other items really are measuring the latent construct more effectively. Alternatively they may be important items that have been well specified, but are essentially single item scales, rather than the multi-item scales the research is attempting to develop. In this case the only alternative is to develop further items to better measure this poorly represented construct if the benefits of multi-item scales are to be realised.

Because of the problems inherent in selecting the number of factors purely on the basis of statistical considerations (i.e. the Kaiser criterion), it has been suggested that researchers need to compare different solutions. For example, Hair et al. (1998) suggest the use of the Kaiser criterion to approximate the number of factors, then running the analysis with one less factors, with one more factor and with two more factors, to compare the interpretability of results. More specifically, “by examining a number of different factor structures derived from several trial solutions, the researcher can compare and contrast to arrive at the best representation of the data. An exact quantitative basis for deciding the numbers of factors to extract has not been developed” (Hair et al., 1998, p. 103).

A further issue concerns the actual scale items that are used to describe the latent construct. To aid the selection of items, factor loadings (essentially the correlation between the item and the corresponding factor) are generally used. The squared factor loadings indicate the proportion of the variance in an original item that is explained by a particular factor. In common with most statistical techniques, the significance of a factor loading is dependent on sample size. However, the evolution of marketing research has resulted in the situation where a factor loading of 0.3 is generally assumed sufficient without consideration of the sample size (see Chae and Hill, 2000; Stewart and McAuley, 2000). Of course, it is actually the case that the lower the sample size, the higher the loading must be to indicate an item significantly loading on a given factor. More specifically, guidelines in Hair et al. (1998) suggest that with samples of 350 or more, the traditionally-used factor loading of 0.3 or greater is significant. However, with samples of 200 a factor loading of 0.4 or greater is needed for the same level of significance, while samples of 100 require loadings of 0.55 or greater for this level of significance. Thus, researchers would be advised to take sample size into account when evaluating the factor loadings of individual items.

Additionally, difficulties of interpretation can arise where an item loads significantly on more than one factor (which has been termed “cross-loading”). Again, the normal practice within marketing research has been to eject such variables since they do not tap only a single construct. However, what significant cross-loading may imply is that the two factors are, to some degree at least, correlated with each other. As will be discussed subsequently, this is hardly surprising in many marketing situations where the items are all related to some degree. Thus, blanket ejection of cross-loading items may again detract from the overall goal of producing robust measures of the underlying, latent constructs.

*Interpretation of results through factor rotation*

Related to the issue of cross-loading is that of factor rotation. While it is beyond the scope of this paper to give a full reiteration of the relevant theory, factor rotation is simply a mathematical attempt to simplify the interpretation of the factor loadings of
individual variables (Sharma, 1996). From the first introduction of classical psychometry into marketing (e.g. Churchill, 1979; Peter, 1979), researchers have generally had the opportunity to choose between two main methods of factor rotation, oblique and orthogonal (see Cattell, 1978).

The main difference between the two methods is that orthogonal rotations constrain the “rotation axes” to be at right angles to each other (i.e. factors are uncorrelated with each other), whereas oblique rotations allow the axes to vary in angle (i.e. factors may be correlated with each other). Despite the fact that the choice is available, it is clearly evident that orthogonal rotations are the rule within marketing research (see Stewart, 1981). However, it could be argued that the popularity of orthogonal rotation methods may stem mainly from non-theoretical reasons, such as convenience and precedent.

Without detailing the geometrics of the issue, factor rotation is performed in order to allow the researcher to interpret the structure of the data more meaningfully, i.e. which items load on which latent constructs (Sharma, 1996). This is helpful since, when there are two or more factors, the variance in each item which is attributable to each common factor can infinitely vary (Sharma, 1996). Rotation can help the researcher to identify the most theoretically plausible factor structure. Thus, a rotated factor solution is by no means the only structure, but merely the one which maximises “simple structure” (i.e. items loading highly on only one factor, thus minimising the incidence of cross-loading).

Since there are an infinite number of possible factor solutions obtainable by various rotations, the theoretical assumptions of the competing factor rotation methods are critical. In the present context the assumption of factor uncorrelation (i.e. orthogonal rotation) or correlation (i.e. oblique rotation) is particularly important. Specifically, a number of psychometricians have argued that uncorrelated factors are hardly ever likely to be the case (e.g. Cattell, 1978; Cattell and Kline, 1977; Hair et al., 1998; Kline, 2000). In terms of pure practicality, this is perhaps best argued by Cattell (1978, p. 104, emphasis in the original) when he states: “it makes sense for factors to be correlated rather than represented artificially in rigid orthogonality, because influences in the real world do get correlated”, and further “we should not expect influences in a common universe to remain mutually uninfluenced and uncorrelated” (Cattell, 1978, p. 128).

Notwithstanding the “common-sense” logic of Cattell’s quotes, it is still possible to argue that a certain set of constructs should in fact be uncorrelated in the real world. However, in statistical terms there is still reason to argue that any uncorrelated set of constructs in the real world will in fact be correlated in any sample taken (Cattell, 1978). Furthermore, even if factors were, by some chance, uncorrelated in a sample (although this would naturally be unknown), one should still use an oblique rotation for a very simple reason. Namely, if the situation is such that it is appropriate to rotate using orthogonal axes, then an oblique rotation will use orthogonal axes since the angle between the axes is unconstrained (see Kline, 2000).

Therefore, there appears little reason in theory to use an orthogonal rotation, which naturally begs the question as to why orthogonal rotation is so prevalent in marketing, and indeed seems to also be so within other disciplines (Stewart, 1981; Sharma, 1996). One reason may be that of pure statistical complexity (which may also be a reason for the use of the more mathematically elegant PCA rather than FA). At the time of his writing, Cattell (1978) implied that the physical effort and computational complication of hand-rotating oblique factor structures could be behind the prevalence of orthogonal
rotations. Furthermore, it is also clear that orthogonal rotational methods were available in computer form before oblique methods (Cattell, 1978). This may explain the lack of explicit consideration given to oblique methods in the early years of marketing psychometrics (e.g. Churchill, 1979). However, today we have no such excuses to fall back on, and it would seem that marketers have done themselves a significant disservice by relying on simple precedent in choosing to use orthogonal rotation, rather than examining the options in more detail.

A further reason why we have continued to use orthogonal rotations is that they may offer statistical advantages which go beyond mere simplicity. Specifically, the creation of correlated factors by use of oblique rotation has potential to cause substantial problems in later model testing due to multicollinearity of the predictor variables. With the prevalence of multiple regression methods (and now structural equation modelling) in marketing research, it seems possible that many more informed researchers have avoided the use of oblique rotations when the latter methods may have in fact been more theoretically appropriate.

Thus the rotation decision seems to boil down to a trade-off between theoretical rigour (which would suggest oblique rotation) and statistical simplicity (which would suggest orthogonal rotation). As a result, it seems that researchers would be advised to reverse the standard procedure, and instead beginning with oblique rotations by default, and only using orthogonal rotations when they were appropriate or necessary. In most cases, there is no theoretical reason to suggest uncorrelated factors in any situation in which FA is used, thus oblique rotations should be employed in the first instance. As Cattell (1978, p. 128) argues; “one does not need to confound logical, ideational independence with statistical independence in the domain of empirical observation ... [t]he sad fact for research is that constraint to an artificial orthogonality destroys both the correctness of the pattern discovered and its constancy from one research to another”. As a result, marketing theories may become at worst sample-specific, and not generalisable to other contexts, a situation which seems to go against the very idea of scientific measure development in marketing.

Summary
Thus, when subjected to a more rigorous scrutiny, it can be seen that many of the unquestioned doctrines regarding EFA in marketing research should be viewed not as laws, but instead as suggested starting points for more detailed and thoughtful analysis. Furthermore, it would also seem that some of the procedures we commonly utilise are potentially conceptually incorrect, regardless of the fact that their practical differences may be minor in many cases (see Stewart, 1981). Thus, from the above a number of issues can be identified that researchers need to be aware of and take into account when using FA to purify scales:

- The conceptual choice between PCA and FA must be explicitly addressed. PCA may be suitable for some purposes such as creating indices (see Sharma, 1996), while FA is more appropriate in the case of scales designed to reflect latent constructs.

- Items with low communality should not be automatically ejected from the analysis. Further work should be undertaken to determine whether the low communality was due to error variance or specific variance. In the case of the latter the evidence may point towards the need to revisit earlier development
stages, to develop further items to measure the unique latent construct represented by the single item.

- The number of factors selected to represent the domain of interest should not be slavishly identified using the Kaiser criterion. Rather, alternative solutions and dimensionalities should be examined and compared for interpretability. Where prior theory or empirical research suggests a particular factor structure, confirmatory approaches are to be preferred over EFA (Gerbing and Hamilton, 1996).

- In rotating factor solutions, explicit attention should be given to whether orthogonal or oblique rotation fits best with the theory being tested. Whatever rotation method is selected (e.g., VARIMAX, PROMAX, OBLIMIN etc.) should be justified conceptually rather than just through citation of precedent.

Conclusions, caveats, and limitations
This paper has focused squarely on a single aspect of marketing research, measure development, and further on a single aspect of that—the item purification process in the classical psychometric paradigm. As a result, we recognize that one of the primary limitations of this paper is that very focus. In particular, we offer no specific insights into the wider context of measure development (e.g., formative measures), or regarding other stages of the classical paradigm (such as item generation and data collection). However, we do feel that the general points argued in this paper do have a wider relevance to scientific marketing research than just classical reflective scale development. That resonance concerns our generally high regard for the technical, operational and procedural aspects of research methodology, to the exclusion of conceptual foundations of those methods. With the exception of a minority of scholars, it is repeatedly the case that new methodologies are introduced into marketing, relatively unquestioningly accepted, then rigidly applied until the next methodological advance. This has been observed with, for example, classical psychometrics (e.g., Churchill 1979), CFA methodology (e.g., Gerbing and Anderson, 1988), and also with structural equation modelling (e.g., Bagozzi and Yi, 1988). We strongly believe that scientific marketing scholarship should value a critical conceptual approach to methodology, not an operational or procedural one.

However, we recognize that, in writing such a paper as this, we run the risk of influencing and increasing further the dogma surrounding classical measure development, rather than reducing it. We have no wish that the ideas and concepts presented herein replace those “rules” already associated with the classical paradigm with simply another set of rules to abide by. Our arguments should be viewed as supplements to the tenets of psychometric marketing measure development, not replacements.

In conclusion, the discussion of factor rotation above presents in a microcosm the entire point of this paper. Simply put, marketing researchers who subscribe to the idea of generalisable theories need reliable, valid, and ultimately generalisable measure development methods with which to measure the constructs which our theories are built around. However, we have tried to show that many of the methods which have been used in the past to create our most vital construct measures (and indeed are still being used today) may actually detract from the validity and generalisability of our
measures. Even worse, the reliance on rigid adherence to precedent has led to a situation where our use of such methods is hardly questioned within marketing.

Nevertheless, this paper should not be seen as a diatribe against the classical model of measurement theory, but more as a defence of the theory itself. More specifically, we have tried to show that it is not the theory which is at fault, but that sometimes the link between the theory and techniques we use is broken, which leads us to employ inappropriate methods which do not create measures in line with classical theory (although those same methods may be appropriate in other measure development situations).

What this paper should be seen as a criticism of is dogma. For too long have marketing academics (ourselves included – mea culpa!) been content to merely quote precedence in the use of techniques to develop measures, without spending time examining what these techniques were actually doing. Furthermore, even though newer methods have arrived to largely solve a number of the technical problems we discuss (such as assessing dimensionality), including CFA (e.g. Gerbing and Anderson, 1988; Gerbing and Hamilton, 1996), it should not become the case that these newer methods are used in the same way as the old, i.e. simply by referring to precedence. Instead, we strongly believe that it is part of our remit as scholars to understand the link between our theories and our techniques, and while it is unnecessary to continue to restate such things in research outputs (which is why the use of precedence indeed evolved), we at least need to understand them ourselves. In other words, we should be able to interpret results of our measure development tools, not simply report them.

Note
1. In fact, it can be argued that a number of well-established marketing scales are excellent examples of this (R).

References


Further reading


Appendix. Glossary of terms


*Composite reliability formula*. A more recent approach to estimating reliability of a multi-item scale seemingly first articulated by Jöreskog (1971). This formula does not assume equal item reliabilities, and appears to have been introduced into mainstream marketing in the 1980s (see Fornell and Larcker, 1981; Gerbing and Anderson, 1988).
Confirmatory factor analysis (CFA). A factor analysis where the precise structure of the model is hypothesised, based on some theory. Thus true CFA is a theory testing approach (Sharma, 1996). However, CFA in marketing appears to have come to mean any factor analysis using covariance structure analysis techniques.

Eigenvalue (also known as the Latent Root). The column sum of the squared loadings for any given factor. In essence, it is the amount of total variance in the data set accounted for by the factor (Hair et al., 1998).

Exploratory factor analysis (EFA). A factor analysis where no structure is pre-specified, and the data are used to help reveal or suggest the structure of the model. EFA is therefore a theory building approach (Sharma, 1996). However, within marketing EFA appears to have come to mean only analyses conducted using “traditional” factoring techniques such as principal axis or principal components factoring.

Factor loading. The correlations between the original variables and the extracted factors, vital to understanding the factor itself (Hair et al., 1998).

Factor rotation. The manipulation of the factor axes in order to achieve a simpler and more meaningful solution (Hair et al., 1998).

Formative index. A measure where the observed indicators are assumed to cause a latent construct rather than be caused by (as in reflective scales) it (Bollen, 1989; Diamantopoulos and Winklhofer, 2001). An example of a formative index (sometimes erroneously called a formative scale) is socioeconomic status (SES), which is a combination of four indicators. If any one of these increases, SES also increases, which would not necessarily be accompanied by an increase in the other three (Diamantopoulos and Winklhofer, 2001). Thus the indicators give rise to the construct.

Internal consistency. The extent to which the items in a scale are inter-correlated (DeVellis, 1991).

Kaiser criterion (also known as the Latent Roots criterion). A rule for deciding the number of factors to extract from a factor analysis. The Kaiser criterion states that only factors which explain more variance in the data set than one of the individual items should be retained (DeVellis, 1991; Sharma, 1996). This is usually operationalised by only retaining factors with an eigenvalue of greater than one (see Sharma, 1996).

Latent construct. The underlying phenomenon which a scale (or sometimes an index) is intended to measure (DeVellis, 1991). Latent constructs are not directly observable (Nunnally, 1967), but can be measured (within definable limits of validity) by observing subjects’ responses to (for example) a multi-item scale.

Measurement error. General term for the inaccuracies in measuring a subject’s “true” score on a latent construct, due to the shortcomings of the measuring instrument (Hair et al., 1998).

Multicollinearity. The extent to which one variable in a model can be explained by the other variables (Hair et al., 1998).

Multi-item scales (also called summated scales or composite measures). The combining of several indicators that measure the same (generally latent) construct into a single variable in order to reap the benefits (e.g. increased reliability) of multivariate measurement (Hair et al., 1998).
Oblique factor rotation. A factor rotation in which the factors are extracted in such a way that their axes are free to vary in angle, rather than restricted to orthogonality (Hair et al., 1998). This means that the factors are allowed (but are not necessarily, depending on the angle of the axes which results) to correlate (Kline, 2000).

Orthogonal factor rotation. A factor rotation in which the factors are extracted in such a way that their axes remain at 90 degrees to each other (Hair et al., 1998). This means that the factors are independent of one another and must have a correlation of 0 (Kline, 2000).

PCA. A data reduction technique where the sole aim is to reduce the set of items to a smaller number of factors which explain the maximum amount of variance. Variables are not composed of unique and common variance, and each factor is simply a linear combination of variables (Hair et al., 1998; Sharma, 1996).

Precision estimate. The precision of Cronbach’s (1951) coefficient alpha is measured in terms of the standard error of the intercorrelations between the items in the scale, which is a function of the variance of those intercorrelations. The precision of a perfectly unidimensional scale will be 0. However, as the scale departs from unidimensionality, precision estimates become larger than 0. Thus, the precision estimate is a symptom of multidimensionality, although is not proof of the latter (see Cortina (1993) for a fuller explication of the precision of alpha).

Reflective scales. A measure where the underlying latent construct is assumed to give rise to the observed indicators (Diamantopoulos and Winklhofer, 2001). This model is underpinned by much psychometric theory (see Kline, 2000) and within marketing appears to be the rule rather than the exception, although scholars have argued against its unthinking application (e.g. Bollen, 1989; Bollen and Lennox, 1991; Diamantopoulos and Winklhofer, 2001).

Reliability. The amount to which a variable (or set of variables such as a multi-item scale) is consistent in what it intends to measure. This means that multiple administrations of the measure should return consistent results (Hair et al., 1998). In conceptual terms, if a measure is perfectly reliable, random error is 0. However, systematic error may still exist.

Structural equation modelling. A term used to refer mainly to covariance structure analysis, where hypothesised relationships among latent constructs, and also latent constructs and their observed indicators, are tested using the covariance matrix derived from the observed data (see Diamantopoulos and Siguaw, 2000).

Validity. The extent to which a variable (or set of variables such as a multi-item scale) correctly represents the construct it is designed to measure (Hair et al., 1998). A valid measure should have little systematic as well as little random error.