Research
On the use of construct reliability in MIS research: a meta-analysis

Patrick Y.K. Chau1,*

Department of Information and Systems Management, School of Business and Management, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Received 12 March 1998; accepted 12 October 1998

Abstract

This paper reports results of a meta-analysis of construct reliability measures reported in MIS research. Based on 418 measures from 63 articles published in four major journals, it is observed that although scales developed more recently are no more reliable than those developed in the 1980s, the reliability of most scales reported in the literature is above generally accepted levels. However, for scales used in basic research, more than 40% of them had a reliability lower than the minimal acceptable level, at least 0.80. Scales using interviews for data collection; with more items, obtained from previous studies; and generated from both literature review and interview, were found to have higher reliability. Other research design characteristics, including sample size, type of subjects, scale type, scale format and number of scale points, were found to be insignificant in affecting reliability. Implications of the findings are discussed and guidelines for researchers on the design of their research methodologies with respect to having scales with the level of reliability suitable for the types of research at hand are provided. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Reliability; Scale validation; Meta-analysis

1. Introduction

Management information systems (MIS) research involves the systematic investigation of the development, operation, use, management, and impact of computer-based information systems in organizations [21]. In developing theories in MIS, there is general consensus on the importance of developing valid and reliable measures of constructs studied. For a measure to be valid and possess practical utility, it must be reliable [16]. Conceptually, reliability can be defined as “the degree to which measures are free from error and therefore yield consistent results” ([14], p. 6). In other words, reliability is “essentially an evaluation of measurement accuracy – the extent to which the respondent can answer the same or approximately the same questions the same way each time” ([18], p. 151).

The purpose of the study presented in this research report is threefold. First, to compile reliability coefficients reported in empirical studies in MIS and to compare these coefficients with generally recommended values by individuals such as Nunnally [11]. This gives a basis for an assessment of how
reliable MIS research constructs have been generally. We use the values recommended in Nunnally [11] as a ‘comparison standard’ [16] for two reasons. First, even though these values have been criticized by several researchers (e.g., [13, 16]), we somehow need a basis for comparison. As Nunnally’s numbers are probably the most referenced ones, using them is believed to be a good choice, at least for the comparison purpose. Second, Nunnally’s [11] recommended values, but not those in Nunnally [10] are chosen because they are more recent and higher in magnitude. This provides a more stringent assessment.

The second objective of the study is to determine the effects, if any, of research design on the reliability of scales using meta-analysis [7]. Previous studies in other disciplines such as marketing have found that reliability tends to vary with several research design characteristics such as the number of scale points, the number of items in the scale, and the method of data collection [4, 15, 16]. This study investigates whether or not these or other relationships are apparent from in MIS studies.

The third objective relates to how we should deal with the reliability issue in the scale validation process. With a review on the role of reliability in the scale validation process and the results obtained from the meta-analysis, guidelines on the development of reliable measures in MIS are suggested.

The report proceeds as follows. The role of reliability in validating scales is first discussed. The discussion of three main categories of research design characteristics is followed. The research method is then described. Findings of the meta-analysis are next presented and interpreted. The report concludes with implications and guidelines on the design of research methodologies with respect to having scales with the level of reliability suitable for the types of research at hand.

2. The role of reliability in validating scales

Validating a scale involves more than testing its reliability and should include at least two other tests: content validity and construct validity. Pedhazur and Schmelkin [13] clearly point out that “reliability is a necessary but not sufficient condition for validity” (p. 81). While reliability is essentially an evaluation of measurement accuracy, content validity and construct validity look into the ‘substance’ of the scale itself. In short, content validity refers to an assessment of whether a scale contains appropriate content to be tested. Construct validity determines the extent to which a scale measures the concept it is supposed to measure [2]. Within it, convergent validity and discriminant validity are usually assessed either through multitrait-multimethod (MTMM) techniques or techniques such as principal components or confirmatory factor analyses [18].

Understanding that reliability test is generally the first step in a scale validation process, there are two main issues that we should look into in order to get a more balanced view of the role of reliability in the whole validation process. The two issues are (1) approaches to assess reliability, and (2) ‘standards’ of reliability. They are discussed below.

2.1. Approaches to assess reliability

An intuitive approach of looking at a reliable scale is that it yields consistent results and is free from measurement error. Measurement error can be of two main types: error from the sampling of situational factors accompanying the administration of items, and error from the sampling of content in the instrument [11]. The first type of error can be viewed as the ‘instability’ of the instrument between two administrations. The second type comes from the ‘non-homogeneity’ of the items. Depending upon the ways to address the error, several approaches have been proposed to carry out the estimation to reliability.

To measure the instability or stability of a scale (and thus its reliability), a method called ‘test–retest’ has been proposed and widely used in many studies. In this method, the same scale is given to the same group of people twice with a certain period of time apart. The correlation coefficient obtained is taken as an estimate of the reliability of the scale. Theoretically speaking, the correlation between two sets of observed scores should be one due to the assumption that the underlying unobservable true scores are constant. The correlation is less than perfect however owing to the error occurred in the administration process. The ‘test–retest’ method is criticized for its potential biases due to carry-over effects, i.e., people remember what they responded in the first round and thus inflate the estimates of the reliability being measured in the
second round. Because of this deficiency, it is recommended that this method generally should not be used or be used only with caution [11, 13].

One way to deal with the carry-over effect inherent in the test–retest method is to conduct equivalent forms test (or called alternative or parallel forms test). Instead of repeating the use of the same scale, an ‘equivalent’ test is used in the second test. ‘Equivalent’ here means the items in the two tests are different (to take away the carry-over effect) but equivalent in terms of the contents in the scale. Again, the correlation coefficient obtained between two forms is taken as an estimate of the reliability of the scale. Obviously, a practical problem of using this method is to have two equivalent forms in the first place. Also, how to determine whether the two forms are equivalent is another big issue that restricts the use of this method.

The second main type of error is the non-homogeneity of items in the scale itself. Theoretically speaking, all items within a scale measure a common characteristic (i.e., the unobservable construct) and should be perfectly correlated. The average correlation among items is generally referred as the ‘internal consistency’ of the scale and the correlation coefficient obtained is called internal consistency reliability. The non-perfect correlation coefficient obtained from observed scores is due to the existence of non-homogeneity of items. A big advantage of this type of reliability is that it only requires single administration which can effectively eliminate the carry-over effect in those two-administration methods described above.

There are two main methods to the estimation of internal consistency reliability. The first one is the ‘split-half’ method. This method requires that a given scale be split into half, and treats each half as if it were an equivalent form for the other. Correlation is thus obtained for the estimation of the reliability just like the case in the equivalent forms method. It is this inherent similarity with the equivalent forms test that it has been warned by many researchers on its use to the estimation of reliability. Furthermore, there are two other weaknesses of this approach. First, many scales, including those developed in MIS research, are not very long, or long enough to make such a split. Second, the fact that there are so many different ways to split the items into half and the different correlations thus obtained make people raise the concern of what the reliability is being measured [11].

Another way to determine the internal consistency reliability is the use of Cronbach’s coefficient α which is based on the average correlation among items and the number of items in a scale. Developed by Cronbach in the early 1950s [6], it is by large the most popular use of coefficient for estimation to reliability. There are a number of advantages of coefficient α over other estimation methods. First, it is based on less restrictive assumptions than some other estimation methods such as split-half. Second, it “sets an upper limit to the reliability of tests” ([11], p. 230). In other words, if a scale is found to be of low coefficient α, it is not worthwhile to make other estimates as they would be even lower. Third, it is very easy to compute and in most situations, it provides a good estimate of reliability. Pedhazur and Schmelkin ([13], p. 93) provides an expression of coefficient α as follows:

\[
\alpha = \frac{k}{k-1} \left(1 - \frac{\Sigma \sigma_i^2}{\sigma_x^2}\right)
\]

here \(k\)=the number of items; \(\Sigma \sigma_i^2\)=the sum of the variances of the items; and \(\sigma_x^2\)=the variance of the total score.

Fourth, empirical studies have revealed that no statistically significant differences exist among estimations to reliability using various methods [4, 12]. Given that in single-administration methods, the major source of measurement error is from the sampling of content, it is suggested by Nunnally [11] that “when the domain of content is easily specified, where there is little subjectivity of scoring, and where people tend to vary little over short periods of time, coefficient \(\alpha\) will provide an excellent estimate of reliability. This is the case, for example, for most tests of aptitude and achievement” (p. 232).

The above advantages perhaps explain why the coefficient \(\alpha\) is the most widely used measure for the estimation to reliability of a scale\(^2\).

### 2.2. ‘Standards’ of reliability

Although reporting reliability of the scale used, in a study, has been long emphasized in scale validation

---

\(^2\)For a recent, more detailed discussion of Cronbach’s coefficient \(\alpha\), see Cortina [5].
literature, varying minimum values for the reliability coefficient have been recommended for different types of studies. Though criticized by other researchers, the most ‘authoritative’ guidelines are those of Nunnally [10, 11]. In Nunnally [11], a minimum level of 0.7 for preliminary research, 0.8 for basic research, and 0.9 for applied research was suggested.

There are at least three things that are worth looking into this ‘standards’ matter. First, theoretically speaking, other things being equal, the higher the reliability of a scale the better. However, achieving a higher reliability for a scale probably comes with higher costs and one should understand the costs and benefits associated with it. Some researchers, such as Boyle [3], also warn that a high reliability for a scale may only be due to a high level of item redundancy. Many items in the scale are essentially the same but phrased in several different ways. As in the case of Cronbach’s \( \alpha \), more items generate high reliability. Given that, a highly reliable scale, if measured by the Cronbach’s \( \alpha \) to be higher than 0.90, may only be a ‘misnomer’ ([3], p. 291).

Second, closely related to the first one, the choice of a minimum level of reliability should depend on how the scale will be used. There are again two things here. The first is the type of research being conducted, as, for example, suggested by Nunnally on different levels for preliminary, basic, and applied research. It seems not justified to spend enormous effort to push a reliability coefficient to a very high level (e.g., from 0.7 to 0.9) in the early stages of research. Similarly, in some applied studies where important decisions are made based on specific scores generated from the scale, a very high reliability (e.g., 0.9 or above) is required. The second is the type of construct measured by the scale. Depending on the number of independent and dependent variables in the theoretical model under investigation, it may be acceptable if there are only a few low reliabilities among many independent variables. Contrarily, low reliabilities may be more problematic for dependent variables.

Third, depending on the type of approach or method being chosen for the estimation of reliability, one should also understand the inherent properties or weaknesses of the method selected. For example, coefficient \( \alpha \) effectively measures the internal consistency of a scale, but not homogeneity. In other words, the \( \alpha \) can be high even when the instrument is not uni-dimensional, i.e., a multi-factor measure. Also, the formula for calculating the coefficient \( \alpha \) lends itself to be of a higher score when there are more items in the scale. In other words, the \( \alpha \) is sensitive to the length of the scale.

This completes the discussion on the role of reliability in validating scales. The key thing here is that researchers should not restrict themselves to any guidelines provided by any ‘authoritative’ source and should make the decision based on what maximum amount of error can be tolerated, given the specific circumstances of the research study [13].

3. Research design characteristics

Given the importance of reliability in the validation process, many studies have been conducted to look into factors affecting reliability. In particular, the study of the possible influence of research design characteristics on the reliability of measures has a long history. As early as 1928, Symonds [20] proposed a list of twenty-five factors that may influence the reliability of a measure. At least half of these factors are related to research design. Broadly speaking, research design choices can be grouped into three main categories: sampling characteristics, scale characteristics, and scale development procedures [4].

Sampling characteristics represent the nature of the sample employed in the study. They include sample size, type of subjects, and method of data collection. It is generally assumed that these sampling issues, though perhaps having important concerns for sampling errors, have little direct impact on measurement errors or estimates of scale quality. Previous research has found a weak, though significant, impact of these variables on reliability coefficients [4, 16].

Scale characteristics represent the nature of the scales used in the research. These characteristics include the number of items, type of scale, number of scale points, and scale format. Measurement theory strongly supports the notion that scale characteristics affect the reliability coefficients obtained. Empirical evidence is found [4, 15, 16].

Scale development procedures are the procedures used to generate items and develop scales. These

---

3For a detailed discussion of the three main categories of research design characteristics, see [4] and [5].
procedures include source of scale and procedures used to generate scale items. Empirical evidence suggests little impact of these procedures on reliability coefficients [4, 15].

4. Research method

To establish a database for the analysis, we went to the home page of the MIS Survey Research Section edited by Peter Newsted, Sid Huff and Malcolm Munro within the ISWorld Net (http://www.ucalgary/~newsted/surveys.html). This includes a ‘sampler’ page that contains 188 constructs from 43 articles published in MIS Quarterly (MISQ), Management Science (MS), Communications of ACM (CACM), and Information Systems Research (ISR) between 1983 and 1992. As noted by the Section Editors in the home page, the articles were selected because some significant effort had been made in their studies to determine both the instrument’s reliability and validity. Tracing the articles that contain the scales, the characteristics of the actual measurements were obtained. Seven articles were excluded from this study because details of the scales used were not described in the papers. Articles published in the above four journals between 1993 and 1995 were also examined to pick up more scales and their corresponding reliability coefficients. Another twenty articles were obtained and a total of 418 reliability coefficients were collected from these two sources. Information on nine research design characteristics was sought in addition to the values of the reliability coefficients of the scales. These nine characteristics included sample size, type of subjects, data collection method, number of items, scale format, number of scale points, source of scale, and item generation procedures. They were selected because (1) they had been posited as influencing the size of a reliability coefficient and were used in similar meta-studies, and (2) at least 80% of the papers examined had reported these characteristics. Several other research design characteristics, such as difficulty of items in terms of the number of words in the items, though investigated in other similar research, were excluded from this study because they did not meet the second criterion. Table 1 indicates the source of reliability coefficients analyzed in the study.

5. Findings and discussion

5.1. Distribution of reliability coefficients

Fig. 1 depicts the distribution of reliability coefficients. The coefficients ranged from 0.27 to 0.99 with a mean of 0.81 and a standard deviation of 0.12. In general, the majority of reported reliability coefficients surpass the minimal standards recommended by Nunnally [11]. Eighty-four percent of the observed coefficients were 0.70 or greater, 62 percent were 0.80 or greater, and 29 percent were 0.90 or greater.

<table>
<thead>
<tr>
<th>Source</th>
<th>Time period covered</th>
<th>Number of articles</th>
<th>Number of reliability coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIS Quarterly</td>
<td>1983–1995</td>
<td>34</td>
<td>228</td>
</tr>
<tr>
<td>Management science</td>
<td>1983–1995</td>
<td>13</td>
<td>70</td>
</tr>
<tr>
<td>Communications of ACM</td>
<td>1983–1995</td>
<td>7</td>
<td>36</td>
</tr>
<tr>
<td>Information systems research</td>
<td>1990–1995</td>
<td>9</td>
<td>84</td>
</tr>
</tbody>
</table>

Fig. 1. Distribution of reliability.
coefficients were, on the average, higher than those repeated in a similar meta-study of marketing research [16]. In that meta-analysis, the author found that the mean value of reported coefficients was 0.77. Seventy-five percent of the reported coefficients were 0.70 or greater, 49 percent were 0.80 or greater, and 14 percent were 0.90 or greater.

Sixteen percent of the reliability coefficients included in our analysis were below 0.70, Nunnally’s minimal, acceptable level of reliability. Tracing these 68 coefficients reveals that 29 were from studies published in 1993 or later. This seems to suggest that although, on the average, the reported reliability coefficients in MIS studies were no worse than those in other disciplines such as marketing, the call for more explicit attention to investigation of reliability is still an important issue that needs to be addressed.

A closer look of the 418 reliability coefficients further reveals the urgent need for more attention to reliability. Table 2 shows the distribution of the coefficients grouped by type of research (preliminary or basic). For preliminary research, 12% did not surpass the 0.70 minimal acceptable level. For basic research, which requires a 0.80 or higher reliability coefficient, the percentage jumped up to 41%. This clearly indicates the ‘insufficient quality’ of these scales and thus the conclusions drawn from them to be problematic.

Table 2 contains descriptive statistics of the reliability coefficients grouped by source reveals that the mean coefficients observed in different journals were significantly different from each other. Coefficients reported in studies published in MISQ and ISR tended to be higher than those reported in the other two journals. This may be because these latter journals, especially CACM, publish a large proportion of papers that are relatively more application-focused than theory-oriented than vice versa. Application-focused papers, in general, are less ‘theory demanding’. Nevertheless, the differences, though significant, are not substantive.

Table 4 contains descriptive statistics of the reliability coefficients according to year of publication. More recently published studies reported reliability coefficients in their papers more often than studies published in the 1980s. This may reflect that more MIS papers were published in the four journals studied and/or among those MIS articles, more and more of them were empirical studies which contained constructs and variables in the testing. Analysis of variance, however, reveals that in terms of mean value, there was no significant difference among different years. This suggests that scales developed more recently are no more reliable than those developed in the 1980s.

In terms of type of research, studies that reported the development of a scale for a particular construct exhibited a higher average of reliability coefficients than those studies that used a number of scales to test a research model. This perhaps reflects the nature of these two types of studies. In a study reporting the history/process of the development of a new scale, searching for high reliability should be one important objective. During the process of developing and refining a scale, the reliability may be quite low initially. However, the process should not stop until a certain level of reliability has been achieved. A suggested level may be 0.80 as it can be ‘used’ by at least two of three types of research studies categorized by Nunnally [11]. In studies using scales for testing a research

<table>
<thead>
<tr>
<th>Acceptable level</th>
<th>Preliminary</th>
<th>Basic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>11 (12%)</td>
<td>135 (41%)</td>
</tr>
<tr>
<td>At or above</td>
<td>78 (88%)</td>
<td>194 (59%)</td>
</tr>
</tbody>
</table>

* Number of coefficient (% of coefficient in that type of research).

<table>
<thead>
<tr>
<th>Source</th>
<th>Sample size, N</th>
<th>Mean</th>
<th>SD</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIS Quarterly</td>
<td>228</td>
<td>0.824</td>
<td>0.108</td>
<td>F=5.50</td>
</tr>
<tr>
<td>Management science</td>
<td>70</td>
<td>0.781</td>
<td>0.118</td>
<td></td>
</tr>
<tr>
<td>Communications of ACM</td>
<td>36</td>
<td>0.760</td>
<td>0.177</td>
<td></td>
</tr>
<tr>
<td>Information systems research</td>
<td>84</td>
<td>0.827</td>
<td>0.091</td>
<td>p=0.001</td>
</tr>
</tbody>
</table>
model, as argued elsewhere in this report, the fore-

going process may still be valid for a model with quite 
a few variables in which some of them have relatively 
lower reliability. Table 5 reports the analysis of var-

iance of type of research on reliability.

5.2. Research design characteristics

Table 6 reports key results of statistical tests of the 
possible impact of research design characteristics on 
reliability coefficients and a summary comparison of 
Peterson’s [16] meta-analysis and present research 
findings. In general, the present study concurs with 
what have been found in Peterson [16].

In contrast to the expectation based on statistical 
theory, the relationship between sample size and 
reliability was not statistically significant. This finding 
suggests that having a larger sample size does not 
necessarily lead to a higher reliability. A plausible 
explanation to this is that in those studies with smaller 
sample sizes, the researchers tended to pay more 
attention to the design and development of the scale 
and other factors that may affect its reliability. This 
attention compensated the small sample size in terms 
of reliability.

Type of subjects was also not a significant factor 
affecting reliability. No statistically significant differ-
ence was found between studies using students as 
subjects and studies using business persons. This 
finding thus supports the use of students as surrogates 
of business persons as subjects in MIS studies.

The data collection method was found to be a 
statistically significant factor in reliability. Studies 
that collected data through interviews or personal 
administration obtained a significantly higher average 
of reliability than those that collected data by self-
administration or mail survey. Assuming that inter-
viewer’s bias is not an issue here, this finding suggests 
that MIS researchers should consider the trade-off 
between the two data collection methods. Interviews 
are usually more costly in terms of time and effort, but 
produce scales with a higher reliability.

In regard to sample characteristics, Table 5 reveals 
that, with the exception of the number of items, no 
statistically significant relationships were found 
between each of the three remaining scale character-
istics (scale type, scale format, and number of scale 
points) and reliability. The significant relationship 
between the number of scale items and reliability 
suggests that MIS researchers should consider avoid-
ing scales with only a few items. However, one should 
also remember that high reliability of a long instru-
ment is just a mere fact or characteristic of the 
Cronbach’s coefficient \( \alpha \). Also, using a particular type 
or format of scale does not necessarily lead to higher 
reliability. Similarly, whether the scale is five-point, 
seven-point, or other does not significantly affect 
reliability.

Both characteristics in the category of scale de-
velopment procedures were found to have statistically 
significant impacts on reliability. The source of the 
scale was a significant factor influencing reliability. 
Scales obtained from previous studies tended to have 
higher reliability than those newly developed. Newly 
developed scales in general have not gone through 
as many empirical tests as those that have been
Table 6
Relationship between reliability and research design characteristics

<table>
<thead>
<tr>
<th>Research design characteristic</th>
<th>Descriptive statistics (mean, SD, N)</th>
<th>F-Value</th>
<th>p-Value</th>
<th>Significant?</th>
<th>Peterson’s result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampling characteristic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;=100</td>
<td>0.82, 0.12, 105</td>
<td>2.03</td>
<td>0.11</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>101–200</td>
<td>0.79, 0.14, 109</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>201–300</td>
<td>0.82, 0.10, 103</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;300</td>
<td>0.82, 0.10, 101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>students</td>
<td>0.84, 0.12, 58</td>
<td>2.87</td>
<td>0.09</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>business persons</td>
<td>0.81, 0.12, 360</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data collection method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>self/mail</td>
<td>0.81, 0.12, 379</td>
<td>8.79</td>
<td>0.00</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>interview/personal administration</td>
<td>0.86, 0.09, 39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Scale characteristic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of items</td>
<td></td>
<td>4.00</td>
<td>0.00</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>two</td>
<td>0.75, 0.12, 54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>three</td>
<td>0.75, 0.10, 59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>four</td>
<td>0.79, 0.13, 76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>five</td>
<td>0.81, 0.12, 57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>six</td>
<td>0.85, 0.10, 57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seven</td>
<td>0.82, 0.09, 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eight</td>
<td>0.88, 0.07, 31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nine</td>
<td>0.87, 0.11, 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ten</td>
<td>0.87, 0.09, 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eleven or more</td>
<td>0.91, 0.10, 43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of scale</td>
<td></td>
<td>2.14</td>
<td>0.10</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Likert</td>
<td>0.83, 0.11, 281</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>semantic differential</td>
<td>0.84, 0.11, 26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>others</td>
<td>0.82, 0.047, 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale format</td>
<td></td>
<td>0.18</td>
<td>0.67</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>only endpoint labeled</td>
<td>0.84, 0.10, 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>endpoints labeled with numerical/verbal values on inner categories</td>
<td>0.83, 0.11, 283</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of scale points</td>
<td></td>
<td>2.14</td>
<td>0.10</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>five</td>
<td>0.83, 0.10, 158</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seven</td>
<td>0.82, 0.11, 137</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eight</td>
<td>0.80, 0.12, 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ten</td>
<td>0.94, 0.02, 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>others</td>
<td>0.82, 0.04, 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Scale development procedure</strong></td>
<td></td>
<td>4.27</td>
<td>0.02</td>
<td>Yes</td>
<td>Not studied</td>
</tr>
<tr>
<td>Source of scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>borrowed, unmodified</td>
<td>0.84, 0.11, 75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>borrowed, modified</td>
<td>0.82, 0.10, 187</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>specifically developed</td>
<td>0.79, 0.13, 135</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item generation procedure</td>
<td></td>
<td>5.51</td>
<td>0.00</td>
<td>Yes</td>
<td>Not studied</td>
</tr>
<tr>
<td>literature review</td>
<td>0.83, 0.10, 219</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interview</td>
<td>0.78, 0.15, 49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>both of the above</td>
<td>0.80, 0.11, 119</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
developed and/or modified in previous studies. This finding suggests the importance of refining and validating a new scale through empirical studies.

Analogous to the relationship between the source of scale and reliability, the relationship between the item generation procedure and reliability was also significant. Items generated from interviews only tended to have lower reliability than those generated from literature review or both. This again can be explained similarly to the relationship between the source of scale and reliability. Items from literature review have gone through at least one empirical test while those from interviews may have not been empirically tested.

6. Guidelines for researchers and conclusions

A number of recent articles have called for explicit attention to investigate the reliability and validity of scales/measures used in MIS research [19, 21]. This study looked into the reliability issue. From a meta-analysis of the reliability of 418 scales from 63 articles published in four major journals, three major findings were obtained. First, papers published more recently reported reliability of the scales used in the studies more often than those published in the 1980s. Second, the reliability of most scales used in MIS studies was above generally accepted levels although scales developed more recently were no better than those developed in the 1980s. Third, several research characteristics, including data collection method, number of items in the scale, source of scale, and item generation procedure, were found to have significant impact on reliability. Scales using interview for data collection, with more items borrowed from previous studies and generated from both literature review and interview, exhibited higher reliability.

Based on the above findings, certain guidelines for researchers can be suggested and they are summarized in Table 7. These guidelines are in a point-by-point fashion hoping that they can be served as a useful reference for researchers in designing their research methodologies with respect to having scales with the level of reliability suitable for the types of research at hand.

The findings of the study also have a number of implications. First, a call for explicit attention to investigation and improvement of reliability of scales used in MIS research is still necessary. The development of reliable scales represents a useful starting point for improving the quality of MIS research. Results of the current study suggest that not too much improvement on developing reliable scales for MIS studies has been made in the past 13 years. Fortunately, the results of the current study also suggest that there are ways to increase the reliability of a scale with low reliability initially to an acceptable level. Researchers should plan carefully and consider adopting those techniques/procedures that have been found to have a positive impact on reliability in this study.

Second, comparing the ‘significant’ characteristics with the ‘non-significant’ characteristics, such as type of scale, scale format and number of scale points, a clear difference between these two groups of characteristics is that the former, in general, requires a lot more effort to generate scales than the latter, and researchers can more easily control the characteristics of the latter group. For instance, collecting data and generating items for a scale by interviews requires more time and money than that by only mail survey or literature review. On the other hand, the effort required to make a five-point or seven-point scale, or with endpoints labeled only or with numerical/verbal values on inner categories, should not have any real differences. This finding suggests the need for securing more effort, in terms of time and money, on the development or generation of reliable scales for MIS studies, especially when the study is a piece of applied, not preliminary or exploratory, research.

Of course, the findings of this study do not suggest that those non-significant characteristics are unimportant issues. Reliability is just one of the many issues that need to be addressed in instrument validation. Others, such as construct validity and statistical conclusion validity, are equally important [18]. For instance, sample size is certainly an important issue when a study plans to use structural equation modeling or LISREL for the examination of construct validity and data analysis [2, 17]. This kind of statistical technique requires a minimal sample size of 150 or even 200 [1]. Similarly, whether the scale should be of Likert-type or semantic differential depends on the assumptions made by the researchers about the relationship between the respondent’s underlying attitude and the responses that will be given to the individual items that make up the scale [8].
Fourth, the fact that significant relationships were found between reliability and a number of research design characteristics including data collection method, source of scale and item generation procedure, may indicate the lack of adequate skills of some MIS researchers in these areas. Measurement theory has suggested and empirical tests in other discipline such as Marketing have found insignificant relationships between reliability and these areas [4, 16]. More emphasis and efforts should be put in to ensure that MIS researchers see the importance and possess adequate levels of these skills.

Fifth, in reporting reliability coefficients, researchers should describe the characteristics of the scales and the procedures of generating them in as much detail as possible. This should include the background of the scale, appropriate references, previous reliability estimates (if any), and the information of the research design characteristics investigated in the current study. This information may help other researchers to assess

In conclusion, this research report has documented the magnitudes of reliability coefficients obtained in MIS research in four major journals that have published MIS papers over the past 13 years. The findings of the meta-analysis suggest that MIS researchers should pay more attention to a number of research design characteristics in order to achieve a scale with a high reliability. A limitation of the current study is that, compared to similar studies in other fields, such as [16] in marketing, the sample size was much smaller, 418 versus 4286. While the sample size was considered to be large enough for the type of statistical analysis conducted in the study, another meta-analysis, perhaps after five years, should be conducted again to re-examine the issues being investigated here. By then, there should be a much larger sample size of reliability coefficients, and cross-validation using two separate sample splits of the data.
should be used in order to reduce the likelihood that the results obtained are due to chance.

Newsted and Huff have pioneered the work on improving the validity of instruments used in MIS research by creating a catalog of data acquisition instruments for MIS researchers [9] and developing a home page of survey instruments on the Internet. This paper continues their effort and has gone one step further by conducting a meta-analysis on the use of reliability in MIS research. It is hoped that this report will not just provide useful information for MIS researchers when developing and evaluating instruments, but also encourage more papers to be published with more information on the scales used in the studies. More importantly, it is hoped that this study will stimulate MIS researchers to put more effort into validation of instruments used in their studies.

References


Patrick Y.K. Chau is an Associate Professor of Information and Systems management at the Hong Kong University of Science and Technology. He received his Ph.D. in business administration from the Richard Ivey School of Business (formerly Western Business School), The University of Western Ontario, Canada. His research interests include issues related to IS/IT adoption and implementation, information presentation and model visualization, and decision support systems. He has papers published in major information systems journals including Management Information Systems Quarterly, Journal of Management Information Systems, Decision Sciences, Information and Management, Journal of Organizational Computing and Electronic Commerce, European Journal of Information Systems, and INFOR. From March 1999 onwards he will serve as an Associate Professor at the School of Business, the University of Hong Kong.