Randomizing Survey Question Order
Vs. Grouping Questions by Construct:
An Empirical Test of the Impact
On Apparent Reliabilities and Links to Related Constructs

Dale L. Goodhue
University of Georgia
dgoodhue@terry.uga.edu

Eleanor T. Loiacono
Worcester Polytechnic Institute
eloiacon@wpi.edu

Abstract

There has been a quiet debate about the proper way to design a questionnaire when multiple questions measure each construct. Two general arguments favor intermixing the questions randomly with those of other constructs; one argument favors grouping together questions that measure a single construct. This study tests these arguments using two different versions of a web quality instrument. One version has questions grouped and labeled, and the other has questions randomly intermixed. Six hundred and three undergraduate students filled out questionnaires, half in each condition. Several different analyses suggest that the grouped question treatment has higher calculated Cronbach's alpha reliabilities than intermixed questions, but is actually less reliable.

1. Introduction

In organizational research that requires measurement, the practice of using multiple items for each underlying construct is fairly well established [1]. Using multiple items increases the reliability of our measures (the ratio of true variance to observed variance [3], and also allows us to estimate how reliable the measure is. The first is important since unreliable measures can mask or cloud true relationships between constructs. The second is important since we need to know how reliable measures are before we can interpret the existence (or lack thereof) of empirically indicated links between constructs.

There has been a quiet debate about the advisability of different approaches to placing those multiple items on a questionnaire. Budd [2] and Goodhue [7] have suggested that items for all constructs to be measured should be randomly placed on the pages so that no two items measuring the same construct are adjacent; Davis and Venkatesh [5] have suggested that grouping together the questions for each construct may improve the quality of the measures.

Budd [2] and Davis and Venkatesh [5] have both conducted empirical studies supporting their point of view. Budd used as his context the Theory of Reasoned Action (TRA) [6] and measured (among other things): attitudes, beliefs about consequences, and intent to perform three different behaviors (brushing teeth at least 3 times a day, smoking cigarettes, and exercising for 20 minutes or more each day). Davis and Venkatesh used as their context the Technology Acceptance Model (TAM) [4], a specialization of TRA to the domain of information technology use.

We suspect that the Technology Acceptance Model may not be the best context for testing the impact of question order on measurement quality. This is because the constructs for TAM (ease of use and usefulness) are unusually well defined, very salient and immediately understandable to respondents, and the measures have been tested and refined by a very large number of studies. The place where we need the strongest attention to
measurement issues, and where the results from this type of research will be the most helpful, is in the development of new measures, not in the repeated testing of established measures. Therefore we will use a newly developed measure as the context for our investigation of the impact of question ordering on measurement quality.

In this paper we conduct a careful test of the impact of two different placement schemes for 36 questions measuring 12 constructs of internet web site quality. In one placement scheme, all questions on the same construct are placed together under a label for the construct. In the other placement scheme, the 36 questions are randomly placed, with the proviso that no two questions measuring the same construct are adjacent. 603 students filled out the questionnaires, about half in each condition.

2. Conceptual Background

2.1. Reliability as a Key Indicator of Measurement Quality

Reliability is a key aspect of measurement validity. Reliability is the extent to which measurements are repeatable [10] or contain a relatively high proportion of true score and relatively low proportion of random error [3]. In plainer language, reliability tells us the degree to which a measure is clouded by random error. We often estimate the reliability of a multi-item measure by Cronbach's alpha, which uses the inter-question correlations as the basis for estimating reliability. But Cronbach's alpha as a measure of reliability is based on several assumptions that are not always met in organizational research.

Cronbach's alpha assumes parallel measures with uncorrelated error terms. More precisely, this means that the scores on the i-th question must be representable by $S_i = T + e_i$, with T being the underlying true score shared by all the parallel questions, $e_i$ being random error for the i-th question, and with the random error terms $e_i$ uncorrelated between questions.

Some researchers, citing Novick and Lewis [9], have taken comfort in the thought that Cronbach's alpha is a lower bound for reliability, even if the above assumptions are not strictly true. However, Novick and Lewis only considered the impact of relaxing one critical assumption for parallel measures. Specifically they asked, suppose that different questions had true scores that were linearly related but not equal. They found that under these circumstances, reliability would be at least as good as that indicated by Cronbach's alpha. However, Novick and Lewis did not consider the impact of relaxing the requirement that random errors are uncorrelated between questions. If error terms across questions are correlated (as they probably are if similar questions are adjacent), then Cronbach's alpha would overstate reliability, rather than giving a lower bound.

There are three general arguments that have been proposed for favoring grouping or intermixing questions on a survey. (1) Grouping artificially inflates reported reliability. (2) Grouping encourages hypothesis guessing and inflates the strength of links to related constructs. (3) Intermixing confuses respondents, lowering both the reliability and the strength of links to related constructs. We will discuss each of these in turn.

2.2. Argument 1: Grouping Artificially Inflates Reported Reliability

One argument for randomizing question ordering is that not doing so artificially inflates the calculated reliability of measures [2, 7]. Goodhue [7] has argued that if several questions for the same construct are adjacent, individuals will tend to use anchoring and adjustment in deciding upon their answers, perhaps answering the first question independently of the others, but making only minor adjustments in answering the remaining questions. Thus the error terms for adjacent questions should be correlated, and this would inflate the apparent reliability, without increasing the actual reliability. This suggests that putting multiple questions for a construct adjacent (as is often done in MIS research) probably artificially inflates calculated Cronbach's alphas.

Budd [2] tested this assertion in a study of 188 undergraduate students, 97 of which completed a randomized questionnaire measuring constructs from the Theory of Reasoned Action [6] pertaining to three different actions. The other 91 students completed a grouped version of the same questions. Budd found that merely placing questions for each construct adjacent increased Cronbach's alpha from about .5 to about .8. While Budd concluded that the grouped questions produced an inflated estimate of reliability, we recognize that there are other possible interpretations of the higher reliabilities. We defer the discussion of possible interpretations until later, and merely hypothesized the following:

H1: placing multiple questions on each construct adjacent to each other (rather than intermixing questions on different questions
randomly) will result in higher (apparent) reliability as calculated by Cronbach's alpha.

2.3. Argument 2: Grouping Encourages Hypothesis Guessing and Biases Answers in Their Favor.

Budd [2] suggests that respondents may recognize the intended relationships between constructs, and will try to make their answers consistent with those relationships in an effort not to be inconsistent. He points out from his own empirical work that when asked to rate the honesty of persons filling out a questionnaire measuring constructs from the theory of reasoned action [6], people tended to attribute the most dishonesty to those questionnaires that did not display consistency between attitudes, social norms, and behaviors.

Thus, Budd argues, individuals are aware of the expected relationships between these constructs, and in an effort to be consistent, will tend to bias their answers to conform with the expected model. This will be far easier to do, Budd argues, if questions are grouped by construct, than if they are randomized throughout the questionnaire. Grouping questions should therefore lead to inflated estimates of links between constructs. In his study Budd found that intermixing the questions resulted in much weaker links between constructs from the theory of reasoned action. Again, we defer the discussion of other possible interpretations until later, and hypothesize the following:

H2: placing multiple questions on each construct adjacent to each other (rather than intermixing questions on different questions randomly) will result in stronger (apparent) links to other constructs, where those links are conceptually justified and "guessable" by respondents.

2.4. Argument 3: Intermixing Confuses the Respondent

Davis and Venkatesh [5] propose an alternate interpretation for the results found by Budd. They argue that by constantly shifting from construct to construct and back again, respondents are forced to repeatedly retrieve cognitions from long term memory, and suffer "output interference" or confusion and frustration which impedes their ability to accurately retrieve all but the early constructs. Davis and Venkatesh conducted a series of studies of 708 undergraduate students and found that grouping versus intermixing the questions had no statistically significant impact on reliability or the links between constructs. However, they did find that comments from respondents after filling out the questionnaire suggested that those with the randomized placement found the questionnaire more confusing and frustrating.

Based on this evidence, Davis and Venkatesh argued that randomizing the questions may actually reduce the quality of the measures. More specifically, if respondents with intermixed questions were more confused, it is quite plausible they would demonstrate lower reliability, and weaker links between constructs.

Note that this interpretation is quite different from Budd's above, but is still consistent with Hypothesis 1 and 2. Thus Hypotheses 1 and 2 by themselves will not be able to distinguish between these two interpretations. Both interpretations predict that grouped questions will have higher reliability and stronger links between constructs. Budd argues that the intermixed questions reflect the "true" picture, and that grouping questions artificially inflates correlations between adjacent questions and facilitates hypothesis guessing. Davis and Venkatesh argue that the grouped questions reflect the "true" picture, and that intermixing questions introduces extra confusion and frustration, interjecting more error into all questions.

To sort out what is really happening we need a way to determine which approach produces the most reliable measure. But we cannot rely on Cronbach's alpha as a way of estimating reliability since if Budd is right, Cronbach's alpha overestimates reliability for grouped questions. Below we describe such an approach.

2.5. How Much More Reliable Is A Composite Measure Than Its Individual Questions?

The most important motivation for asking multiple questions is that the composite measure is presumed to be more reliable than any of the individual questions. It is important to clarify the reason why multiple questions should improve reliability. The presumption is that if distinct questions really can be represented by \( S_i = T + \epsilon_i \), as described above, and if the error terms are uncorrelated, then when several distinct questions are combined, the error terms will tend to cancel out, and the result will be a measure with a higher proportion of true score to error. On the other hand, if the error terms are correlated, then the error terms will not tend to cancel out, and the
combined score will not be more reliable than the individual questions.

As a way of illustrating this more concretely, consider that we have three questions measuring say, "response time of a web site", and two questions measuring "intent to revisit the web site". Consider further that we have theoretical reasons for assuming that response time is significantly correlated with intent to revisit. Since we presume that response time and intent are correlated, one indication of reliability is the degree of correlation between individual questions across the two constructs. In this case we have six such correlations (three questions for response time, multiplied by two questions for intent). The average of these six correlations is in some sense an indication of the reliability of the set of three questions for response time.

Now consider the composite measure created by averaging the three questions for response time, and the composite measure for intent created by averaging its two questions. If the error terms for each of the questions are uncorrelated, then averaging the questions for each construct should result in the uncorrelated errors canceling out, at least to some extent. Therefore the reliability of the composite measures should be higher than the average reliability of the individual questions.

If on the other hand, the error terms are correlated, then averaging will not cause errors to cancel out, and we would not expect the composite measure to be more reliable than the average reliability of the individual questions. In this case we would not expect the composite measure of response time to have a higher correlation with intent than the average of the individual question correlations.

Applying this line of thinking to the dilemma discussed in this paper, it is likely that for both grouped questions and intermixed questions on a survey, the error terms will be at least somewhat uncorrelated. Therefore we would expect that the composite measures will be more highly correlated with related constructs than the average of the individual question correlations. However, following Budd's and Goodhue's argument about anchoring and adjusting for grouped questions, we believe that the error terms for grouped questions will be relatively more correlated, that those error terms will tend to cancel out to a smaller degree, and that therefore we will see a smaller improvement in reliability as we move from an average of the individual question correlations with the related construct to the composite measure correlation with the related construct. This leads to the following hypothesis:

H3: Assuming multiple questions for measuring any given construct, as we move from A) an average of individual question correlations with a related construct to B) the composite measure correlation with that same related construct, improvement will be greater for intermixed questions than for grouped questions.

This last hypothesis discriminates between the interpretation of Budd and Goodhue on the one hand, and Davis and Venkatesh on the other.

3. Methodology

We tested the above hypotheses with an experiment involving 603 undergraduate students at a southeastern university. Students were enrolled in or had just completed an introductory course in MIS that included studying and accessing the web. All students were given the option of completing the experimental exercise or completing a short report, either of which would give extra credit. Approximately 1200 questionnaires were distributed, and 603 were returned with complete data, for a response rate of about 50%. Students were asked to visit one of 12 randomly assigned web sites, investigate the web site with the presumption that they were planning a trip or looking for a gift for a friend (depending upon which of the 12 web sites was used), and then to fill out a questionnaire about the web site quality. The questionnaire had been developed as part of the doctoral dissertation of one of the authors, and measured 12 components of web site quality with three questions each, or 36 questions total. In addition, after the body of the questionnaire, students answered two questions asking whether they were likely to make a purchase from this web site, and whether they were likely to revisit this web site.

A random half of the students were given questionnaires with the 36 questions randomly intermixed with no two questions on the same construct adjacent, and the other half were given questionnaires with the 36 questions grouped
together, with a label identifying each set of three questions. This second treatment is presumed to be an extreme condition likely to generate anchoring and adjustment\(^2\).

Both versions of the questionnaire measured 12 constructs of web site quality, from a measure of web site quality developed as the doctoral dissertation of one of the authors [8]. The 12 constructs are: informational fit to task, interactivity, trust, response time, design, intuitiveness, visual appeal, innovativeness, flow, integrated communications, facilitation of business processes, and ability to substitute for other means of interacting. The derivation and justification of these constructs is beyond the scope of the current paper. Of course the quality of measurement of the constructs is quite germane to the current paper, and will be dealt with below. All twelve constructs are assumed to have a causal link with the intent to reuse the web site, along the lines of the theory of reasoned action [6], though we have followed Davis and Venkatesh's [5] lead in excluding attitudes from the model, based on the finding that attitudes do not entirely mediate between beliefs and intent.

4. Experimental Results

4.1. LISREL Measurement Model Analysis

We first tested the fit of the data from the two treatment groups by using LISREL to test their fit to the measurement model. Table 1 shows the essential fit statistics for the two treatment groups. All paths from constructs to their questions were highly significant with t statistics of greater than 8.0. Thus both treatment approaches result in good model fits as tested by LISREL, though the grouped question treatment resulted in a better fit.

4.2. Cronbach's Alpha Reliabilities

Cronbach's alpha reliabilities were calculated for the two treatment groups: intermixed and grouped, as shown below in Table 2. Calculated reliabilities were greater in the grouped treatment for 11 of the twelve constructs. We know of no test of whether difference in the alpha values on any given construct are statistically significant across the two conditions, but we can use a non-parametric test based on the assertion that (a null hypothesis for H1) there were no systematic differences in the alpha values. If this is true, for any given construct the likelihood that the grouped alpha is higher than the intermixed alpha is .5 (i.e. 50%). Then the probability of the grouped treatment having only one construct (out of 12) with reliability greater than the intermixed treatment is .003\(^3\). Thus we would reject the null hypothesis that reliabilities are equal (p<.01). Hypothesis H1 that the grouped treatment would have higher reliabilities is supported.

4.3. Exploratory Factor Analysis of Internal Consistency and Discriminant Validity.

To complete the assessment of measurement validity to the two instruments we then performed an exploratory factor analysis of data for each treatment condition. For both treatments we used an oblique rotation method (Promax) since we expect that these constructs are correlated, though distinct. For the grouped question treatment, we used an eigenvalue cutoff of .71 to produce 12 factors. All questions loaded cleanly on their constructs, with all but 4 loadings in excess of .84. Those four loadings were: .78, .75, .74, and .66. The highest cross loading was .20. This is an excellent fit to the data.

The results were not as strong for the intermixed questions. We used an eigenvalue cutoff of .63 to produce 12 factors. The loadings were similarly high for seven of the twelve constructs. However, questions for "information..."
Table 2. Cronbach’s Alpha Reliabilities for Intermixed and Grouped

<table>
<thead>
<tr>
<th></th>
<th>Intermixed</th>
<th>Grouped</th>
<th>Higher Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>0.86</td>
<td>0.91</td>
<td>Grouped</td>
</tr>
<tr>
<td>Inter</td>
<td>0.80</td>
<td>0.91</td>
<td>Grouped</td>
</tr>
<tr>
<td>Trust</td>
<td>0.90</td>
<td>0.94</td>
<td>Grouped</td>
</tr>
<tr>
<td>Resp</td>
<td>0.88</td>
<td>0.89</td>
<td>Grouped</td>
</tr>
<tr>
<td>Design</td>
<td>0.83</td>
<td>0.90</td>
<td>Grouped</td>
</tr>
<tr>
<td>Intuit</td>
<td>0.79</td>
<td>0.92</td>
<td>Grouped</td>
</tr>
<tr>
<td>Visual</td>
<td>0.93</td>
<td>0.98</td>
<td>Grouped</td>
</tr>
<tr>
<td>Innov</td>
<td>0.87</td>
<td>0.91</td>
<td>Grouped</td>
</tr>
<tr>
<td>Flow</td>
<td>0.81</td>
<td>0.90</td>
<td>Grouped</td>
</tr>
<tr>
<td>IntCom</td>
<td>0.87</td>
<td>0.94</td>
<td>Grouped</td>
</tr>
<tr>
<td>BusProc</td>
<td>0.72</td>
<td>0.91</td>
<td>Grouped</td>
</tr>
<tr>
<td>Subst</td>
<td>0.81</td>
<td>0.81</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Average</td>
<td>0.84</td>
<td>0.91</td>
<td></td>
</tr>
</tbody>
</table>

4.4. Correlations with "Intent to Reuse"

In each version of the questionnaire, two questions were asked for the construct "intent to reuse". (See appendix for the wording.) The correlations between these two questions were, in the intermixed and the grouped treatments respectively: .72 and .78, resulting in Cronbach's alphas of .84 and .88. Table 3 shows the correlations between the 12 constructs of web quality and "intent to reuse".

H2 suggested that individuals facing grouped questions would be able to guess hypotheses and bias their answers to be consistent with those. Thus we expected to see that correlations with "intent to reuse" would be higher for the grouped treatment. This was not the case. As Table 3 shows, for 10 of the 12 constructs, the correlation with "intent" was higher for the intermixed treatment.

Table 3. Correlations Between Web Quality Constructs and "Intent to Reuse".

<table>
<thead>
<tr>
<th></th>
<th>Intermixed</th>
<th>Grouped</th>
<th>Higher Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>0.52279</td>
<td>0.47386</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Inter</td>
<td>0.49393</td>
<td>0.37213</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Trust</td>
<td>0.26422</td>
<td>0.3614</td>
<td>Grouped</td>
</tr>
<tr>
<td>Resp</td>
<td>0.30089</td>
<td>0.17144</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Design</td>
<td>0.36861</td>
<td>0.36008</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Intuit</td>
<td>0.40022</td>
<td>0.31036</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Visual</td>
<td>0.31898</td>
<td>0.35095</td>
<td>Grouped</td>
</tr>
<tr>
<td>Innov</td>
<td>0.48417</td>
<td>0.4397</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Flow</td>
<td>0.52225</td>
<td>0.50033</td>
<td>Intermixed</td>
</tr>
<tr>
<td>IntCom</td>
<td>0.38858</td>
<td>0.35596</td>
<td>Intermixed</td>
</tr>
<tr>
<td>BusProc</td>
<td>0.43414</td>
<td>0.35045</td>
<td>Intermixed</td>
</tr>
<tr>
<td>Subst</td>
<td>0.44181</td>
<td>0.3652</td>
<td>Intermixed</td>
</tr>
</tbody>
</table>

Again using a null hypothesis that about half the constructs in the grouped treatment would have a higher correlation, and using the binomial distribution, we see that the probability of 2 or fewer successes in 12 trials is .019. Thus we reject the null hypothesis that there is no difference between the treatments (p<0.5). H2 is not supported. In fact, the evidence strongly suggests that the intermixed questions have a higher correlation with "intent to reuse", exactly the opposite of H2.
4.5. Testing the Improvement in Reliability Due to Uncorrelated Errors Canceling Out

Table 4 shows a sample of the data used to test Hypothesis 3. This hypothesis predicts that the intermixed treatment will have a greater improvement in the correlation with "intent to revisit" as we shift our view from the average of the individual WebQual question correlations with the intent to reuse questions, to correlations between the composite measures of both constructs. This is presumably because intermixed questions have uncorrelated errors terms that tend to cancel out.

In Table 4 it can be seen that for the intermixed questions on Fit to Task, the average correlation between individual Fit questions and individual Reuse questions is .41 in the grouped condition, and .43 in the intermixed condition. As we move to composite measures rather than individual questions, the improvement is .09 for the intermixed condition, and only .06 for the grouped condition. Thus the direction of the effect is that suggested by Hypothesis 3, though we have no test of the significance of this on a construct by construct basis. However, taking as a null hypothesis that about half the constructs should be improved more in the intermixed condition, we can again use a non parametric test. For 11 out of twelve constructs, this H3 prediction is correct.

Table 4 shows a sample of the data used to test Hypothesis 3. This hypothesis predicts that the intermixed treatment will have a greater improvement in the correlation with "intent to revisit" as we shift our view from the average of the individual WebQual question correlations with the intent to reuse questions, to correlations between the composite measures of both constructs. This is presumably because intermixed questions have uncorrelated errors terms that tend to cancel out.

In Table 4 it can be seen that for the intermixed questions on Fit to Task, the average correlation between individual Fit questions and individual Reuse questions is .41 in the grouped condition, and .43 in the intermixed condition. As we move to composite measures rather than individual questions, the improvement is .09 for the intermixed condition, and only .06 for the grouped condition. Thus the direction of the effect is that suggested by Hypothesis 3, though we have no test of the significance of this on a construct by construct basis. However, taking as a null hypothesis that about half the constructs should be improved more in the intermixed condition, we can again use a non parametric test. For 11 out of twelve constructs, this H3 prediction is correct.

Table 4. Example of Analysis Showing Extent of Improvement in Correlations with "Intent to Reuse" When Moving From an Average of Individual Question Correlations To Composite Measure Correlations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For Intermixed Questions, N=307</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIT1</td>
<td>0.46</td>
<td>0.44</td>
<td>0.43</td>
<td>0.52</td>
<td>0.09</td>
</tr>
<tr>
<td>FIT2</td>
<td>0.38</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIT3</td>
<td>0.43</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>For Grouped Questions, N = 296</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIT1</td>
<td>0.42</td>
<td>0.41</td>
<td>0.41</td>
<td>0.47</td>
<td>0.06</td>
</tr>
<tr>
<td>FIT2</td>
<td>0.37</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIT3</td>
<td>0.41</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Again using the binomial distribution, we can see that the probability of only 1 or 0 successes in 12 trials is .0032. Thus we would reject the null hypothesis that intermixing has no systematic effect (p<.01), and H3 is supported.

5. Discussion and Further Exploratory Analysis.

These results are somewhat surprising. Intermixing questions, as opposed to grouping and labeling them, results in lower "calculated" Cronbach's alpha reliabilities (i.e. an apparently lower quality measure), but even so, the composite measures from intermixed questions are more highly correlated with related variables. This would seem to make untenable Davis and Venkatesh's suggestion that intermixing confuses respondents, resulting in lower quality measures. However, there is still one possible explanation that is consistent with Davis and Venkatesh's suggestion of output overload and confusion. Suppose that respondents were confused by intermixed questions, and therefore fell back on a more general sense of positive or negative overall reactions to the web site. In this case, for the intermixed treatment we would expect to see all 36 questions sharing some variance based on this overall reaction, and this shared variance could also be linked to "intent to reuse". In other words,
intermixing questions would result in an instrument bias linking all questions. Such a bias could explain the higher correlations we observed with "intent to reuse". We can test for this possibility in LISREL.

5.1. Analysis to Test Instrument Bias Due to Confusion.

To do this we use the basic model of 12 constructs with 3 questions each, but add two new constructs: "intent to reuse", and "instrument bias". The "intent to reuse" construct is measured as before by two questions (see appendix), and the "instrument bias" construct is modeled as affecting all of the original 36 questions equally. This allows us to test to see if there is some overarching factor having an impact on all questions.

The result of this is two LISREL models (one for the intermixed treatment and one for the grouped treatment) which fit about as well as the original models, but now show an instrument bias across all questions. This bias exists in both treatments, but is stronger in the intermixed treatment. To see the relative size of the bias, for the intermixed treatment, the path loading for the bias is .30 (p<0.001), compared with the path loadings for the questions of .58 to 1.10. For the grouped treatment, the path loading for the bias is .23 (p<0.05), compared to the path loadings for the questions of .73 to 1.21. However, even when we account for the bias in this way, the correlations between the 12 constructs and "intent to reuse" are still higher for the intermixed treatment than in 10 out of 12 cases. The probability of only 2 or less successes in 12 trials is .019. Therefore we would reject the null hypothesis that there is no difference between the treatments, even after adjusting for a "confusion factor bias" (p<0.02). The difference is still in favor of the intermixed treatment.

5.2. One Further Exploratory Analysis.

To carry our analysis one step further, we will come close to paralleling a compelling portion of the analysis done by Davis and Venkatesh [5]. They looked at the regression path coefficients between ease of use and usefulness, and added an interaction term to capture the degree to which intermixing the questions changed the value of the path coefficient. This is really the central issue for determining whether intermixing improves or degrades reliability.

In testing six interaction terms (three different samples and interaction terms for both ease of use and usefulness in each), they found no significant interaction terms. A power analysis suggested that they had sufficient power to detect the interaction effect if it would have caused a change in R² of more than .05. In Davis and Venkatesh's data, we would conclude that grouping versus intermixing does not cause a change in R² of more than .05.

We tried to parallel that analysis as much as possible with our data. We are not in a position to predict "intent to reuse" in a regression using all 12 web quality constructs, since we know that many of them are correlated and thus multicollinearity would severely cloud our ability to interpret the results. Developing and testing a second order factor model prior to predicting intent is also beyond the scope of this paper. However, we can use each of the 12 factors separately as predictors of "intent to reuse", and in this way approximate the interaction test made by Davis and Venkatesh.

To do this we combined all 603 questionnaire into a single dataset, with a new variable called "intermixed" which is 1 if the questions were intermixed, and 0 if they were grouped. We then ran 12 regressions, each with one of our 12 web quality constructs, the zero-one intermixed variable, and the interaction of the two (intermixed times the web quality construct) as predictors of "intent to reuse". This closely parallels the analysis made by Davis and Venkatesh.

Our findings are not that different from Davis and Venkatesh's. In only one of the 12 regressions did we find a significant (at the .05 level) interaction term. Apparently the impact of intermixing is not great enough to create a statistically significant change in the path coefficient when the interactions are looked at one at a time. However, this data is worth a closer look.

If intermixing had no impact on the path coefficients, we would expect about half of the interaction terms to be positive, and half negative. On the other hand if intermixing had a small but systematic impact on the path coefficients, we would expect a predominance of positive interaction terms. Therefore the null hypothesis would be a probability of .5 that each individual interaction term would be positive. In fact, 11 of the 12 interactions terms were positive. This makes the null hypothesis of no impact untenable at the .01 level. From our data, it is fairly clear that intermixing systematically improves the path coefficients, but only by a small amount. A closer

---

4 Once again, a binomial distribution with 1 success in 12 trials, with a probability of .5.
look at Davis and Venkatesh's data shows exactly the same result. For each of the six interaction terms they tested, the value was positive, indicating a small (but in the aggregate statistically significant at the .05 level) increase in the path coefficient when intermixed questions were used.

6. Conclusion

We are left with the following picture. Grouping questions results in significantly higher Cronbach's alphas (higher by an average of .07 in this case), but it is not tenable to claim that the measures of the constructs are actually of higher reliability. The predicted links between the 12 constructs and "intent" are nearly all lower for the grouped treatment, even when controlling for the possibility of a smaller "confusion factor bias". Our analysis paralleling Davis and Venkatesh's regression with interaction terms gives the same results, as does a reexamination of Davis and Venkatesh's results. We conclude that Cronbach's alpha is artificially inflated when questions are placed adjacent to each other and labeled, and that intermixing questions results in a small but systematic improvement in actual reliability.

Further, we have presented a somewhat unusual but compelling demonstration of the extent to which correlations with related constructs are improved when shifting from an average of individual question correlations to composite measure correlations. Thus we see that the reliability of the constructs seems to be improved more by this shift when using intermixed questions. This is consistent with the contention that when questions are intermixed, error terms are less correlated, and averaging across questions is more likely to cancel out errors. There does seem to be a somewhat larger "confusion factor" when questions are intermixed, but that does not overwhelm the extra advantage of improving reliability by having uncorrelated error terms.

We note that Davis and Venkatesh's TAM questions are very strong and well tested measures, while our WebQual measures are new and just being tested. One interpretation of our findings might be that intermixed questions are needed at least for newly developed measures. However, recall that a reanalysis of Davis and Venkatesh shows exactly the same pattern and significance: the effect was not statistically significant on any individual link, but overall the effect is statistically significant (11 out of 12 interactions were positive in our study, and 6 out of 6 in Davis and Venkatesh’s). Though more replications should be done, the initial evidence (from our study and Davis and Venkatesh’s) is fairly strong that intermixing questions creates measures with slightly higher actual reliability, even though the calculated Cronbach's alpha will be less.

7. References


Appendix:
Questions from WebQual\textsubscript{TM} Survey

(7 point scale, strongly disagree -- strongly agree)

Informational Fit-to-Task
The information on the Web site is pretty much what I need to carry out my tasks.
The Web site adequately meets my information needs.
The information on the Web site is effective.

Interactivity
The Web site allows me to interact with it to receive tailored information.
The Web site has interactive features, which help me accomplish my task.
I can interact with the Web site in order to get information tailored to my specific needs.

Trust
I feel safe in my transactions with the Web site.
I trust the Web site to keep my personal information safe.
I trust the Web site administrators will not misuse my personal information.

Response Time
When I use the Web site there is very little waiting time between my actions and the Web site’s response.
The Web site loads quickly.
The Web site takes long to load.

Design
The display pages within the Web site are easy to read.
The text on the Web site is easy to read.
The Web site labels are easy to understand.

Intuitiveness
Learning to operate the Web site is easy for me.
It would be easy for me to become skillful at using the Web site.
I find the Web site easy to use.

Visual Appeal
The Web site is visually pleasing.
The Web site displays visually pleasing design.
The Web site is visually appealing.

Innovativeness
The Web site is innovative.
The Web site design is innovative.
The Web site is creative.

Flow
I feel happy when I use the Web site.
I feel cheerful when I use the Web site.
I feel sociable when I use the Web site.

Integrated Communications
The Web site projects an image consistent with the company’s image.
The Web site fits with my image of the company.
The Web site’s image matches that of the company.

Business Processes
The Web site allows transactions on-line.
All my business with the company can be completed via the Web site.
Most all business processes can be completed via the Web site.

Substitutability
It is easier to use the Web site to complete my business with the company than it is to telephone, fax, or mail a representative.
The Web site is easier to use than calling an organizational representative agent on the phone.
The Web site is an alternative to calling customer service or sales.

Additional questions for "intent to reuse"

How likely or unlikely would you be to make a purchase from this Web site? (1 "Very unlikely to purchase" to 7 "Very likely to purchase")

How likely or unlikely would you be to revisit this Web site? (1 "Very unlikely to revisit" to 7 "Very likely to revisit")