Summer 6-15-2016

A Q-SORTING METHODOLOGY FOR COMPUTER-ADAPTIVE SURVEYS – STYLE "RESEARCH"

Sahar Sabbaghan
University of Auckland, s.sabbaghan@auckland.ac.nz

Lesley A. Gardner
University of Auckland, l.gardner@auckland.ac.nz

Cecil Chua
University of Auckland, aeh.chua@auckland.ac.nz

A Q-SORTING METHODOLOGY FOR COMPUTER-ADAPTIVE SURVEYS – STYLE "RESEARCH"

Sahar Sabbaghan, Department of ISOM, University of Auckland, Auckland, NZ, s.sabbaghan@auckland.ac.nz

Lesley Gardner, Department of ISOM, University of Auckland, Auckland, NZ, l.gardner@auckland.ac.nz

Cecil Eng Huang Chua, Department of ISOM, University of Auckland, Auckland, NZ, aeh.chua@auckland.ac.nz

Abstract

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. Due to the complexity of CAS, little work has been done on developing methods for validating their construct validity. This paper describes the process of using a variant of Q-sorting to validate a CAS item bank. The method and preliminary results are presented. In addition, lessons learned from this study are discussed.

Key words: Computer-Adaptive Surveys, Construct Validity, Q-sorting.

1 Introduction

One new type of survey in business research is the Computer-Adaptive Survey (CAS). Unlike in a traditional survey, where every question is asked (Hayes, 1992), in a CAS, the previous questions determine the next questions asked of the respondent. The CAS is seldom used in business research, because each respondent must be provided with a computational device to perform the survey. However, the ubiquity of mobile devices means the use of CAS for business research is becoming increasingly feasible.

A CAS is especially useful to identify an issue a respondent affiliates with/rejects most. Examples of domains that CAS is relevant to include evaluating customer satisfaction, and identifying political issues most salient to the respondent. Within this domain, CAS offers certain advantages over traditional surveys. One is that it allows the researcher to drill in to what is the respondent's major issues. In contrast, traditional surveys are more effective when the aim is to obtain an overall view or perception. To illustrate, the typical customer satisfaction survey is either very short (e.g., 6 items) (Hallowell, 1996), or comprehensive (e.g., 138 items) (Feinberg and Johnson, 1998). A short survey can (for example) identify that a customer is dissatisfied with a product or service such as food. But the specific issues with the food are harder to ascertain from Likert-scale questions. In contrast, when faced with a long questionnaire, the respondent is likely to encounter fatigue and quit before providing critical information (Galesic and Bosnjak, 2009; Groves, 2006; Groves et al., 2004; Heerwegh and Loosveldt, 2006; Porter, Whitcomb, and Weitzer, 2004). Of course, open ended questions can address these issues, but the high variation in open ended responses makes open ended questions difficult to analyse (Jackson and Trochim, 2002).With a CAS, only questions most relevant to a respondent's concerns are asked.

To illustrate CAS, let's consider an example instrument to evaluate customer satisfaction of a café experience focusing on identifying areas of least satisfaction. Initially, as shown in Figure 1, the respondent is presented with generic questions about the quality of food, convenience, and the price. This is the starting point. The responses are sorted from highest to lowest. If food is the area that the customer is least satisfied with, CAS would then present the customer with questions about the preparation, portion, and menu choices. Depending on the respondent's further responses, different questions are loaded until finally the questionnaire determines that there are insufficient vegetarian items on the menu.

CAS problems are inherently IS problems. Unlike other surveys, CAS requires that the respondent have access to a computational device, because algorithms in the background select future questions.

Traditional methods for validating surveys do not work for CAS for two reasons. First, CAS is arranged in a hierarchy. Traditional methods assume a "flat" set of items. Second, respondents legitimately only fill in some of the questionnaire items- unfilled questions cannot be treated as non-responses.

This paper presents a technique we have developed to evaluate the construct validity of a CAS. Specifically, we propose a new q-sorting method, which is applied to a CAS on café satisfaction. The results help researchers not only identify quality constructs, but also facilitate diagnoses of problems with constructs.

The paper is constructed in the following manner. Section 2 introduces the related literature, briefly describing CAS and Q-sorting. Following from this, we present a modified Q-sorting approach in section 3. We next demonstrate an example of the use of the Q-sorting approach on a survey in section 4. We then discuss the lessons we have learned through this process in section 5 and finally conclude the paper in section 6.

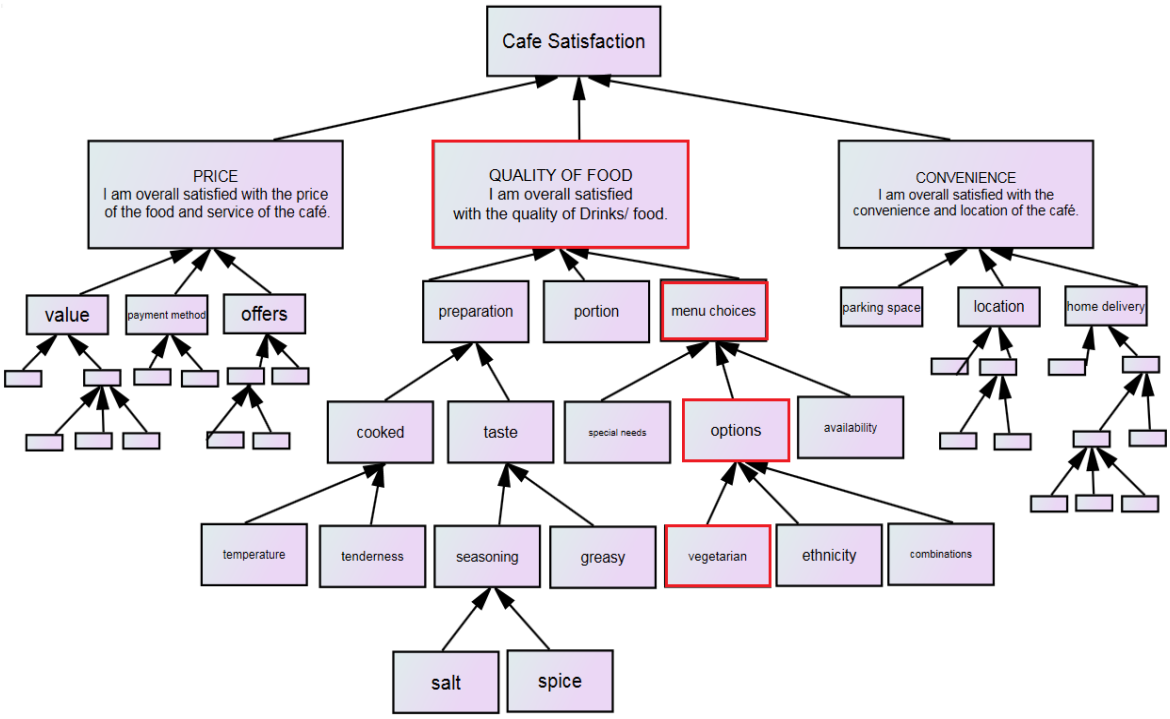


Figure 1. How CAS works for café experience

2 Related Research

There are numerous real-world situations where someone desires to ask a large battery of questions to obtain in-depth information about individuals' perceptions. Examples include surveys on job satisfaction (Chien et al., 2011), or customer satisfaction (Meyer and Schwager, 2007). For example, someone who asks customers about customer satisfaction typically wants to know not only that something is wrong, but precisely what is wrong (Fundin and Elg, 2010; Sampson, 1998; Wisner and Corney, 2001).

A CAS can be thought of as a hybrid of a traditional perception survey and a Computer-Adaptive Test. Computer-Adaptive Tests (CAT), are designed to efficiently assess and evaluate each participant's latent trait by administrating questions which are dynamically assigned based on answers to the questions the participant answered previously (Gershon, 2005; Thompson and Weiss, 2011). For example, the Graduate Management Admission Test (GMAT) asks the respondent to answer language and mathematical questions in increasing order of difficulty (Stricker, Wilder, and Bridgeman, 2006). The next question asked of a respondent depends on whether the previous questions were answered correctly. Similarly, in the Merrell and Tymms test (Merrell and Tymms, 2007), the aim is to understand the reading ability of students to provide better feedback and implement appropriate reading techniques.

While CAT assesses an ability or performance (Hol, Vorst, and Mellenbergh, 2008; Merrell and Tymms, 2007), CAS measures individuals' perceptions. Traditionally, CAT has focused on cognition and behaviors; the goal of the typical CAT is to produce a score evaluating ability or performance on a single or few constructs. The focus of CAT is on individual performance. The goal of CAS is typically to identify one or a few narrowly defined constructs that respondents as a whole have the greatest or least affiliation to. The interest is in determining what the group likes or dislikes. For example, a CAS of café satisfaction would try to determine which of several elements of a café a group of respondents most dislike. The last question asked in a CAT reflects the respondent's ability- thus, the last few questions asked on the GMAT are reflective of the respondent's language or math skill. The last question asked on a CAS on café satisfaction reflects the things the respondent most dislikes about the café. When aggregated, this reveals what the group most dislikes about the café.

These dissimilarities in goals result in structural incongruities between the two kinds of surveys. A CAT will typically contain a large number of questions about one or two "main" constructs having several child constructs. For example, the GMAT measures a respondent's ability on "math" and "verbal" skills. A CAS will typically have more "main" constructs. For example, a CAS on café satisfaction may have five constructs, (1) convenience, (2) service quality, (3) quality of food and drink, (4) price and value, and (5) ambiance.

CAT relies on potentially complicated Item Response Theory (IRT) functions to determine further questions to ask respondents (Embretson and Reise, 2000; Lord, 1980; Thompson and Weiss, 2011; Thorpe and Favia, 2012). CAS, in contrast, uses an adaptive version of branching to arrange the questions. Low or high scores on a set of questions causes the system to retrieve related, but more precise questions.

The question structures also differ. On the GMAT, which is based on IRT, the "correct" answer adds a point to the score, while an incorrect one deducts from 0.25 to 0.20 from one's score. In contrast, items in CAS are more akin to those on traditional psychometric instruments that are designed to "load" on a construct.

Finally, initiation and termination in CAS and CAT function in specific ways. In most cases, respondents taking a particular CAT test all begin in the same way. In contrast, we could have the first 20 subjects taking a CAS begin with generic questions about the café. If we realise that most respondents are indicating issues with the food, the next 20 subjects might begin at a lower level of the hierarchy- on the food-related questions. Similarly, a CAT terminates when the CAT has enough information to perform a diagnosis, either when a fixed number of items have been answered (Babcock and Weiss, 2009; Ho and Dodd, 2012; Shin, Chien, and Way, 2012) or because further questions in the item bank provide no additionally statistically meaningful information (Thompson and Weiss, 2011). In contrast, a CAS terminates either when one fully traverses a set number of branches of the hierarchy, or when the user reaches some threshold for a proxy for fatigue (e.g., user answers a certain number of questions).

CAS is likewise, similar, yet different to a traditional perception survey. Like a traditional perception survey, the questions are generally Likert-style- in contrast to CAT, where the questions tend to have a correct answer. However, unlike a traditional perception survey, the expectation is that most questions will be unanswered by a respondent. Also, unlike a traditional perception survey, a CAS cannot be used to find cause in the sense of there being an independent variable and dependent variable on the survey. In a CAS, there is an implicit dependent variable (e.g., customer dissatisfaction) that the survey does not ask. Instead, the survey attempts to get at the root cause of that dependent variable, i.e., why are customers unhappy with the café?

Like a traditional perception survey, it is important to ensure constructs are orthogonal (i.e., independent) to each other. Like a traditional perception survey, there is one exception where constructs that are children to other constructs in the hierarchy should represent some dimension of the parent construct. Thus, a question about the taste of food, and a question about one's perception of the food can exist in the same survey so long as the former question is a child of the latter.

These differences preclude the use of traditional CAT and perception-based validation techniques on CAS. CAT uses several validation techniques (Borman et al., 2001). One technique is expert judgement- an expert with content expertise reviews the questions (Hambleton and Zaal, 2013). This method has limited applicability for CAS, because expert judgement is typically about determining whether a question corresponds to a single category of questions. While in CAT there is only one or two constructs, in CAS, there are multiple constructs and child constructs and the validity problem is determining which of the many child constructs (if any) a question properly fits in.

Another common technique used in CAT is comparing the scores obtained from the CAT test against results from a well-accepted similar non-CAT test (Hambleton and Zaal, 2013; Huff and Sireci, 2001). This method is not suitable for CAS, because CAS does not produce scores as much as its goal is to identify the child constructs most salient to a particular group of respondents.

Construct validation methods used in "flat" perception surveys also are inapplicable, because one must assess the construct validity of an item not only with respect to other items at the same level, but also with respect to an item's parent and children. In traditional perception surveys, there are relatively few of such items. In CAT, all items have this property. Consider the items "(1) The restaurant offered a variety of menu choices," "(2) There were healthy food options available at the restaurant," and "(3) There was food from different cultures." (2), and (3) should relate to (1), because (1) appears to be their parent- (2) and (3) should therefore correlate somewhat with (1). But, (2), and (3) are orthogonal to each other- cultural variety and health should not have a relationship and hence they should not have a strong correlation with each other. Traditional mechanisms for performing construct validity like structural-equation model-based confirmatory factor analysis (Anderson and Gerbing, 1988; Kline, 1998) or rotating factor scores (Hair, Anderson, Tatham, and Black, 1998) thus have limited applicability.

Traditionally, construct validity in surveys is performed using two methods. The first is factor loading, "which is the correlation between the original variables and the factors, and the key to understanding of the nature of a particular factor" (Hair et al., 1998, p. 89). However, this method is problematic for use with CAS, because factor loading is inappropriate for use with items related as parents and children (Diamantopoulos and Winklhofer, 2001; Jarvis, MacKenzie, and Podsakoff, 2003; Mackenzie, Podsakoff, and Podsakoff, 2011).

The second method to assess construct validity is Q-sorting (Straub and Gefen, 2004). In the typical q-sort test for construct validity, independent raters are provided with a set of cards, where each card contains a single questionnaire item. Raters are then instructed to place the cards into groups, where the groups correspond to the constructs (Block, 1961). In some cases, the number of groups is pre-assigned (Segars and Grover, 1998). In others, grouping is left to the rater (McKeown and Thomas, 1988). Q–sorting may be one of the best methods to assess content and construct validity for constructs with parent-child relationships (Petter, Straub, and Rai, 2007).

The traditional q-sort suffers limitations similar to those of other construct validity tests for managing CAS survey items, notably an inability to manage hierarchy. However, we have found a modification of Q-sorting can be employed. This will be further elaborated in section 3.

3 Modified Q-Sort Approach

Our q-sort technique begins during survey development. The aim of this modified q-sort is to validate a tree hierarchy. Our Q-sort technique has the following steps; (1) Incorporation of distractors (2) Recruitment of raters (3) Sorting process (4) Inter-rater agreement process (5) Evaluation of the inter-rated scores

Step 1: Incorporate duplicate and distractor questions in the questionnaire. The typical CAS test bank can contain hundreds of items. The duplicate and distractor questions are used to assess rater attentiveness during the q-sorting process. It is important to ensure that distractor questions are clearly independent of questions being assessed by the rater. When raters miss the fact that questions are duplicated, or categorize distractor questions along with legitimate ones, it signals lack of rater attention. For example, on a CAS about cafes, one distractor question could be "The education level is sufficient." It could be argued that such questions are unnecessary, because poor inter-rater reliability would serve as an effective proxy. However, inter-rater reliability is also indicative of poor question phrasing. It is necessary to be able to partial out the effect of raters and questions separately.

Step 2: Recruit independent raters. At least two independent raters need to be recruited to q-sort the item bank and explicitly draw the tree hierarchy. They should be blind to both the study design and each other. Raters are also trained to draw a tree hierarchy. They are told that each parent must have at least 2 children in the tree hierarchy. Raters do not sort all questions in one session, rather they are given a period of time (e.g., a week) to finish the q-sorting process. This is because a CAS can have over 500 items which could group into over 50 constructs (including child constructs). This is too heavy a cognitive load for raters to perform in a single session. Instead, raters are given only questions from one

top-level construct at a time. In effect, the CAS is treated as multiple CATs. Like a CAT, the rater examines all items associated with a single construct to assess its validity. One limitation of this approach is the possibility that a question sorted within one top-level construct actually belongs with another top-level construct. This issue can be minimized by ensuring the nomological validity of the questions prior to q-sorting (Straub and Gefen, 2004)In other words, care should be taken that the literature has documented that a question is appropriate for a particular top-level construct and not for others.

Step 3: Sort items into constructs. Raters are told to do two things in their sorting. First, they should sort items into constructs. Second, they should map constructs together in a hierarchy, with constructs concerning higher level concepts linking to lower level concepts.

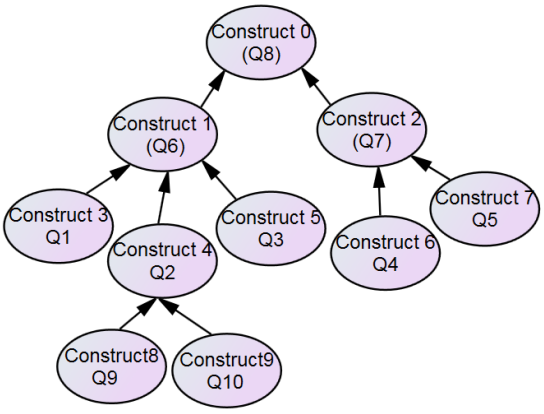
In creating hierarchies, certain rules should be followed:

- A construct can only have child constructs if the construct has two or more children. With only one child, there is no "branching," which obviates the need for a CAS. This rule should only be enforced on a second attempt at q-sorting. In the first attempt, it is useful to allow raters to make mistakes so as to identify potentially problematic questions.
- Branches must be connected to the tree. If both raters identify one branch as unconnected, then that item is either a distractor or does not belong to any identified construct.

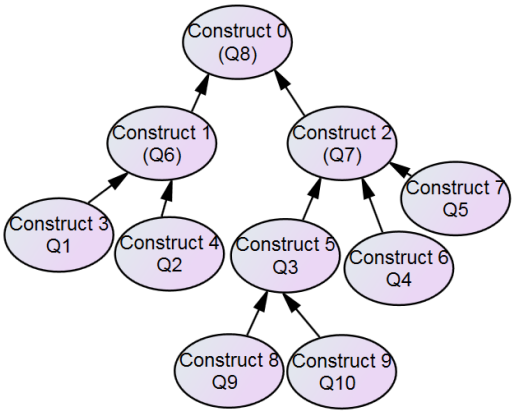
Step 4: The inter-rater agreement of the tree hierarchy needs to be processed. Each rater's mapping is transformed as follows**.** Once the questions are put in the hierarchy, each branch in the hierarchy is assigned a number from 0 to N, where N is the total number of branches. Thus, the first branch is given the number 0, the next branch the number 1, etc. Leaf nodes, i.e., the final sets of questions asked, are treated as branches for this analysis. We then map the question number to the construct number, and map the construct numbers to each other. To illustrate, see Figure 2. For both raters, questions 9 is assigned to construct 8 and question 10 is assigned to construct 9. However, for rater 1, constructs 8 and 9 are mapped to construct 4, while for rater 2 constructs 8 and 9 are mapped to construct 5. In addition, for both raters, Questions 4 and 5 are assigned to constructs 6 and 7. Constructs 6 and 7 are mapped to construct 2. However, for rater 2, construct 5 is also mapped to construct 2, where for rater 1 construct 5 is mapped to construct 1. For both raters questions 6 and 7 are each assigned to constructs 1 and 2. Constructs 1 and 2 for both raters are mapped to construct 0. In this example, several notions need to be highlighted, one is that only one question was assigned to each construct. However, one or more questions can be assigned to each construct. Second, raters may disagree on the assignment of a question(s) for each construct, such as rater 1 may assign question 3 for construct 6 while rater 2 may assign question 3 for construct 5.

Each rater's assignment of a question to a branch is then tabulated, as highlighted in Table 1. Table 1(a) presents how questions are mapped to constructs. Table 1(b) shows how constructs are then mapped to each other in the hierarchy. Hence, as an example, construct 1 and construct 2 for both raters is mapped to construct 0, i.e., construct 1 and construct 2 are children of construct 0. Whereas, for construct 5, rater 1 has mapped it to construct 1 and rater 2 has mapped it to construct 2, i.e., the raters have disagreed and assigned construct 5 as a child of two separate parent constructs. The results in Table 1 correspond to the diagrams in Figure 2.

We then perform an iterative contingency table analysis. In the initial analysis, we compare the raters' mapping of questions to constructs (i.e., Table 1(a)) and obtain both the p-value (significance) and lambda (strength) of this initial mapping. We then move the mapping of questions to constructs from the end of the hierarchy (i.e., the leaf nodes) to their parents and rerun the analysis. Thus, in the second iteration, all questions that rater 1 classified as belonging to construct 8 and 9 are reassigned as belonging to construct 4. All questions that rater 2 classified as belonging to construct 8 and 9 are reassigned to construct 5. We continue to do this until the hierarchy has only two levels. Note that this is effectively the converse of multiple hypothesis testing (Saffer, 1995), in that all tests must be passed for construct validity to be obtained.



(a) Rater 1



(b) Rater 2

Figure 2. Raters' tree diagram example

Questions	Rater 1	Rater 2
question1	Construct 3	Construct 3
question2	Construct 4	Construct 4
question3	Construct 5	Construct 5
question4	Construct 6	Construct 6
question5	Construct 7	Construct 7
question6	Construct 1	Construct 1
question7	Construct 2	Construct 2
question8	Construct 0	Construct 0
question9	Construct 8	Construct 8
Question10	Construct 9	Construct 9

(a) Mapping of question to Construct

Construct	Rater 1	Rater 2
Construct 0	0	0
Construct 1	0	0
Construct 2	0	0
Construct 3	1	1
Construct 4	1	1
Construct 5	1	2
Construct 6	2	2
Construct 7	2	2
Construct 8	4	5
Construct 9	4	5

(b) Mapping of Constructs to Each Other

Table 1. Data in analysis software.

The best possible significance and strength are obtained in the initial analysis. This significance and strength wanes as questions are remapped. This is because significance and strength are based on: (1) the correspondence in mapping from questions to constructs, and (2) the correspondence in mapping between constructs. The initial analysis only takes (1) into account. Later analyses take (2) into account as well. We consider a λ of 0.7 to be a reasonable figure. λ measures the strength of the similarity- it is akin to r in a regression. Note that Cohen (1988) considers r of 0.5 to be a strong correlation, so our threshold is very high. Nevertheless, we will perform future research to investigate acceptable strength thresholds. The inter-rater reliability test we run has to pass all of the tests, not just one of them, for us to consider the hierarchy reliable.

Name	df	χ^2	*p*	Λ
Construct overall for CAS	9	22.44	.04	.857
Construct (remove 1 level)	4	11.600	0.21	.600

Table 2. Contingency Table Test of Inter-Rater Agreement.

As illustrated in Table 2, λ decreases from 0.857 to 0.6. This means that while raters agreed on the mapping of questions to constructs, they disagree on what parent branch the constructs belong to. Our technique not only assesses the validity of the hierarchy, but also diagnoses which problem (mapping of questions to construct or construct to construct) causes problems with inter-rater agreement.

Step 5: Evaluation of the inter-rated scores. When satisfactory levels of significance and strength are achieved, items that raters disagree on need to be reconciled. This is done by assembling the researchers and raters to discuss the discrepancies. Depending on the feedback from raters, questions can be discarded, rewritten, or a final mapping from parent to child construct can be agreed upon.

4 Using the Modified Q-sorting Approach

To determine the effectiveness of our q-sort technique, we developed a CAS which is designed to elicit the problems customers had with cafes. Five overarching constructs were identified: (1) convenience, (2) service quality, (3) quality of food and drink, (4) price and value, and (5) ambiance. Raters were asked to q-sort approximately 176 survey questions with four duplicate questions and distractors. Hence a total of 180 items were given to the raters.

Next, we recruited two independent raters and asked them to sort the items associated with each overarching construct (i.e., questions on convenience are sorted differently from service quality). Raters did not sort all 180 questions at once, rather raters in their own pace, took one overarching construct at a time and sorted them. Hence, their overall cognitive load was reduced. The rater's assignment of a question to a branching is then recorded.

The following were applied to minimize the biases giving raters only a subset of questions could create. First, questions were principally adapted from the café satisfaction literature. Second, care was taken to ensure the overarching constructs were conceptually distinct. It is difficult to imagine individuals confusing a question on the taste of a food item with a question on the service of the waiter, for example. Finally, care was taken to ensure instructions to the rater solely concerned how to develop a hierarchy. No instructions focused on the subject matter of cafes, nor did instructions suggest a certain number of levels, branches (other than there needing to be at least 2 branches of a parent construct) or constructs was correct.

Figure 3 presents an example of how this was coded for the construct "Convenience." A number of concepts should be highlighted in Figure 3. First, both raters identified one construct (called 14 for both raters) as unconnected. This construct contained the distractor questions. Second, the ordering of the numbers is immaterial for the analysis. Thus, that construct 2 which is represented by question 3, is at the same level for rater 1 while it is a child of construct 1 for rater 2 is not important. The analysis is concerned with whether rater 1 mapped the same constructs into construct 3 as rater 2 categorized in construct 4, and whether rater 1 mapped the same questions into construct 4 that rater 2 categorized in construct 5. Finally, because of the nature of the questionnaire, each construct should have at least two questions. Furthermore, constructs that are not "leaf nodes" should branch into two further child constructs. However, Figure 3 has "single" branches, while Rater 2 only assigned a single child construct to construct 6. These are actual rater responses, which could either indicate error in the construct or in this case, the raters made mistakes.

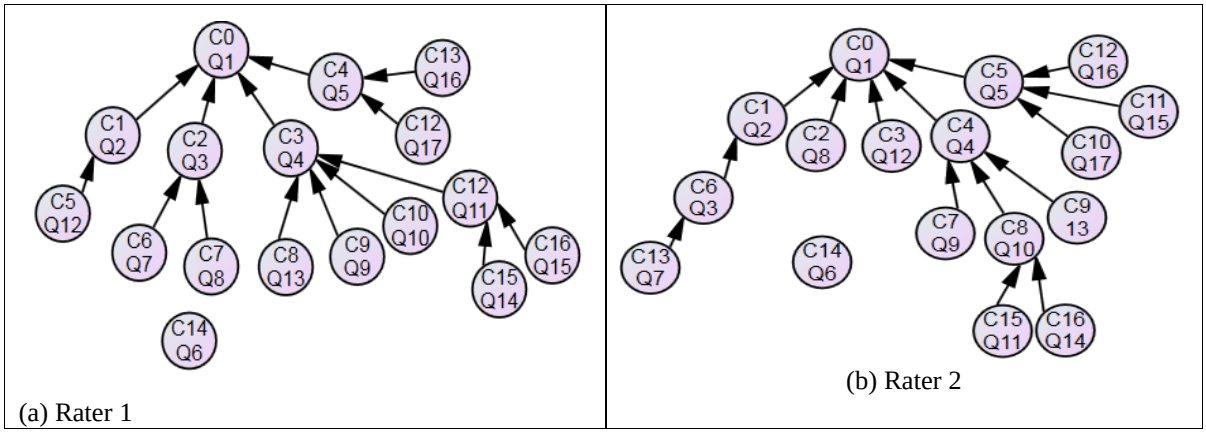


Figure 3. Raters' tree diagram for the construct "Convenience"

We next collapsed the hierarchy by one level. We did this in a bottom up way. For example, referring to Figure 3 (a), for rater 1, C1 was dropped and C5 was assigned as the child of C0 (i.e., C1's parent), C2 was dropped, and C6 and C7 were assigned C0 as a parent (ie., C2's parent). We then reran the analysis again, repeating until there are only two levels in the hierarchy.

Construct	df	χ^2	*p*	λ
Convenience (all)	25	39.067	<.001	.619
Service quality (all)	204	355.153	<.001	.563
Service quality (remove 1 level)	135	348.713	<.001	.506
Service quality (remove 2 level)	45	133.416	<.001	.451
Price and Value (all)	16	54.857	<.001	.704
Environment (all)	144	233.574	<.001	.551
Environment (remove 1 level)	80	173.31	<.001	.507
Environment (remove 2 level)	30	100.60	<.001	.435
Food Quality (all)	100	320.356	<.001	.704
Food Quality (remove 1 level)	24	113.36	<.001	.623
Overall for CAS	1935	4432.545	<.001	.650

Table 3. Contingency Table Test of Inter-Rater Agreement

The current results from the café satisfaction survey are presented as Table 3. As Table 3 demonstrates, λ decreases as one collapses levels. This indicates that while raters think of the questions as belonging to the same group, they disagree on what parent branch the questions belong to. As an example, in the construct Food Quality, when one level is removed λ drops from 0.704 to 0.623. The constructs that are below the threshold, should be evaluated and the problematic questions need to be either removed or edited. In our example, after the first round of Q-sorting, some constructs are still below the threshold. Hence, they need to be assessed. As an example, the construct convenience has a lambda score of 0.619. Since Lambda is still below the threshold, the questions need to be examined. According to Figure 3, raters disagree on 6 questions. These questions need to be assessed. After discussion, questions 11 and 15 were dropped and questions 8 and 10 were rewritten.

5 Lessons Learned

We have been refining the questionnaire over several iterations of Q-sorting, and have learned certain lessons:

Do Not Ask Raters to Sort All Items Simultaneously. CAS typically has many more questions than other surveys. We discovered that this is too cognitively burdensome. Instead, allow raters to sort at their own pace and time. Also raters demonstrated signs of cognitive exhaustion. For example, only one rater found the three distractor questions inserted as a check. Other signs of exhaustion can include accidently dropping questions in the q-sorting process and not using all the questions properly. Also, the statistical results of the Q-sorting were poor for those who failed to identify the distractors.

Ensure Minimum Language Proficiency. There are characteristics of CAS that make certain properties of language particularly salient. Specifically, the hierarchical layout of CAS questions means raters must understand words that clue a reader into whether a question is more or less specific. We found individuals with poor command of the English language fail to comprehend such hierarchy-related words as "overall" or "generally." Also, research has demonstrated that different cultures view hierarchy differently (Gentner and Goldin-meadow, 2003). In our first attempt, we did not specify language proficiency as a factor for choosing our raters, hence in the Q-sorting process, the raters who had very different first languages, failed to identify the key words for hierarchy-related words. As a result, the outcome was not very strong.

Train Raters to Draw the Tree Diagram. The nature of the questionnaire required each construct that has children to have at least two child constructs. It is necessary prior to asking for a q-sort that raters be given examples of tree diagrams from other domains. Without such illustrations, raters tended to perform traditional q-sorts. In this study, we allowed raters to make mistakes, however, it is more helpful and less time consuming if the raters are trained in developing tree diagrams.

Provide Questions In Different Formats. The ideal situation for drawing a hierarchy would be to use digital whiteboards or large screens, where raters can visualize the overall tree hierarchy, save their work, and to be able to move items around. Saving work is important, because raters will often want to restore their work from a prior point. However, in circumstances where these devices are not available, raters can receive the questions in both paper and electronic format. Raters move the paper, then re-represent what they have in electronic format. They save the electronic version, and then continue with the paper.

6 Conclusion

This study presented a variant of Q-sorting designed to evaluate construct validity of survey items in CAS. As CAS items have certain characteristics, such as being multi-dimensional and containing constructs with parent-child relationships, traditional methods are not suitable. In our variant, hierarchies that independent raters develop are transformed into a quantitative form, and that quantitative form is tested to determine the inter-rater reliability of the individual branches in the hierarchy. The hierarchies are then successively transformed to test if they branch in the same way. Dummy questions are inserted as a check on raters.

Thus far, our Q-sorting validation technique has been shown to work. It has successfully identified not only that there are problems with the survey instrument we designed, but also what those problems are. For example, a steadily decreasing λ result in our analysis suggests raters are not mapping constructs to the same parents.

Our current technique employs lambda, the strength of association in a contingency table (Everitt, 1992) as a measure of strength instead of using a traditional measure of inter-rater correspondence like Kappa (Fleiss, Levin, and Cho Paik, 2013) or the intraclass correlation (ICC) (McGraw and Wong, 1996). In other work we have done, we have found that a combination of lambda and kappa provides better

insights into the correspondence of raters. Notably, kappa is more sensitive to situations where raters place a single construct under different parent constructs, while lambda is more sensitive to situations where entire branches are placed on different construct. While Kappa misdiagnoses, lambda identifies this issue.

We are continuing to develop suitable threshold values using multiple measures in the same way that Hu and Bentler (Hu and Bentler, 1998) proposed a combination of CFI, SRMR and RMSEA as a goodness of fit for structural equation models.

In addition, we are refining our technique to better accommodate the heavy cognitive load on raters. Work in this area is ongoing.

References

Anderson, J., and Gerbing, D. (1988). "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach". *Psychological Bulletin 103*(3), 411–423. doi:10.1037/0033-2909.103.3.411

Babcock, B., and Weiss, D. J. (2009). Termination Criteria in Computerized Adaptive Tests : Variable-Length CATs Are Not Biased. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Block, J. (1961). "An introduction to the Q-sort method of personality description". *The Q-Sort Method in Personality Assessment and Psychiatric Research* 3–26. doi:10.2307/1419315

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., and Drasgow, F. (2001). "An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales." *Journal of Applied Psychology*. American Psychological Association. doi:10.1037/0021-9010.86.5.965

Chien, T.-W., Lai, W.-P., Lu, C.-W., Wang, W.-C., Chen, S.-C., Wang, H.-Y., and Su, S.-B. (2011). Web-based computer adaptive assessment of individual perceptions of job satisfaction for hospital workplace employees. *BMC Medical Research Methodology 11*(1), 47. doi:10.1186/1471-2288-11-47

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. *Statistical Power Analysis for the Behavioral Sciences* (Vol. 2nd). doi:10.1234/12345678

Diamantopoulos, A., and Winklhofer, M. (2001). Formative Indicators : Scale. *Journal of Marketing Research 38*(2), 269–277.

Embretson, S. E., and Reise, S. P. (2000). "Item response theory for psychologists." *Quality of Life Research 4*(3), 1– 371. doi:10.1023/B:QURE.0000021503.45367.f2

Everitt, B. S. (1992). *The analysis of contingency tables*. *Monographs on statistics and applied probability* (Vol. 45). Chapman and Hall/CRC.

Feinberg, S., and Johnson, P. Y. (1998). Designing and developing surveys on WWW sites. *Proceedings of the 16th Annual International Conference on Computer Documentation - SIGDOC '98*, 38–42. doi:10.1145/296336.296349

Fleiss, J., Levin, B., and Cho Paik, M. (2013). *Statistical Methods for Rates and Proportions*. *John Wiley & Sons*.

Fundin, A., and Elg, M. (2010). "Continuous learning using dissatisfaction feedback in new product development contexts." *International Journal of Quality & Reliability Management 27*(8), 860–877. doi:10.1108/02656711011075080

Galesic, M., and Bosnjak, M. (2009). "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly 73*(2), 349–360. doi:10.1093/poq/nfp031

Gentner, D., and Goldin-meadow, S. (2003). Lannguage in mind: Advances in the study of language and Thought. In *Language in Mind* (pp. 3–15). Cambridge, MA: MIT Press.

Gershon, R. C. (2005). "Computer adaptive testing." *Journal of Applied Measurement*.

Groves, R. M. (2006). "Nonresponse Rates and Nonresponse Bias in the Household Surveys." *Public Opinion Quarterly 70*(5), 646–675.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. (J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, … T. W. Smith, Eds.)*Statistics* (Vol. 2nd). Wiley-Interscience. doi:10.2307/1504821

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate Data Analysis*. (J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, Eds.)*International Journal of Pharmaceutics* (Vol. 1). Prentice Hall. doi:10.1016/j.ijpharm.2011.02.019

Hallowell, R. (1996). "The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study." *International Journal of Service Industry Management 7*(4), 27–42. doi:10.1108/09564239610129931

Hambleton, R. K., and Zaal, J. N. (2013). *Advances in educational and psychological testing: Theory and applications*. *Springer Science & Business Media*. doi:10.1007/s13398-014-0173-7.2

Hayes, B. E. (1992). *Measuring Customer Satisfaction*. Milwaukee, WI: ASQC Quality Press.

Heerwegh, D., & Loosveldt, G. (2006). An Experimental Study on the Effects of Personalization , Survey Length Statements , Progress Indicators , and Survey Sponsor Logos in Web Surveys. *Journal of Official Statistics 22*(2), 191–210.

Ho, T.-H., and Dodd, B. G. (2012). "Item Selection and Ability Estimation Procedures for a Mixed-Format Adaptive Test." *Applied Measurement in Education*, *25*(4), 305–326. doi:10.1080/08957347.2012.714686

Hol, a. M., Vorst, H. C. M., and Mellenbergh, G. J. (2008). "Computerized Adaptive Testing of Personality Traits." *Zeitschrift Für Psychologie / Journal of Psychology 216*(1), 12–21. doi:10.1027/0044-3409.216.1.12

Hu, L., and Bentler, P. M. (1998). "Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification." *Psychological Methods 3*(4), 424–453. doi:10.1037/1082-989X.3.4.424

Huff, K. L., and Sireci, S. G. (2001). "Validity Issues in Computer-Based Testing." *Educational Measurement, Issues and Practice 20*(3), 16–25. doi:10.1111/j.1745-3992.2001.tb00066.x

Jackson, K. M., and Trochim, W. M. K. (2002). "Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses." *Organizational Research Methods 5*(4), 307–336. doi:10.1177/109442802237114

Jarvis, C. B., MacKenzie, S. B., and Podsakoff, P. M. (2003). "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research." *Journal of Consumer Research 30*(September 2003), 199–218. doi:10.1086/376806

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling* (Third.). New York: The Guilford Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. *Applied Psychological Measurement* (Vol. 5). Hillsdale, N.J.: L. Erlbaum Associates.

Mackenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. (2011). "Construct Measurement and Validation Procedures in Mis and Behavioral Research : Integrating New and Existing Techniques." *MIS Quarterly 35*(2), 293–334.

McGraw, K. O., and Wong, S. P. (1996). "Forming inferences about some intraclass correlation coefficients." *Psychological Methods 1*(1), 30–46. doi:10.1037/1082-989X.1.1.30

McKeown, B. F., and Thomas, D. B. (1988). *Q Methodology (Quantitative Applications in the Social Sciences series, Vol. 66)*. Newbury Park, CA: Sage.

Merrell, C., and Tymms, P. (2007). "Identifying reading problems with computer-adaptive assessments." *Journal of Computer Assisted Learning 23*, 27–35. doi:10.1111/j.1365-2729.2007.00196.x

Meyer, C., and Schwager, A. (2007). Understanding customer exprience. *Harvard Business Review 85*. doi:10.1108/00242539410067746

Petter, S., Straub, D., and Rai, A. (2007). "Specifying Formative Constructs in Information Systems Research." *MIS Quarterly 31*(4), 623–656.

Porter, S. R., Whitcomb, M. E., and Weitzer, W. H. (2004). "Multiple surveys of students and survey fatigue." *New Directions for Institutional Research*, *2004*(121), 63–73. doi:10.1002/ir.101

Saffer, J. P. (1995). "Multiple Hypothesis Testing." *Annual Review of Psychology 46*, 561–584. doi:0.1146/annurev.ps.46.020195.003021

Sampson, S. E. (1998). "Gathering customer feedback via the Internet: instruments and prospects." *Industrial Management & Data Systems 98*(2), 71–82. Retrieved from http://www.emeraldinsight.com/journals.htm?articleid=849897&show=abstract

Segars, A. H., and Grover, V. (1998). "Strategic Information Systems Planning Success: An Investigation of the Construct and Its Measurement." *MIS Quarterly 22*(2), 139. doi:10.2307/249393

Shin, C. D., Chien, Y., and Way, W. D. (2012). *A Comparison of Three Content Balancing Methods for Fixed and Variable Length Computerized Adaptive Tests*.

Straub, D., and Gefen, D. (2004). "Validation Guidelines for IS Positivist Research." *Communications of the Association for Information Systems 13*, 380–427.

Stricker, L. J., Wilder, G. Z., and Bridgeman, B. (2006). "Test Takers' Attitudes and Beliefs About the Graduate Management Admission Test." *International Journal of Testing 6*(3), 255–268. doi:10.1207/s15327574ijt0603_3

Thompson, N., and Weiss, D. (2011). "A framework for the development of computerized adaptive tests." *Practical Assessment, Research & Education 16*(1). Retrieved from http://www.pareonline.net/pdf/v16n1.pdf

Thorpe, G. L., and Favia, A. (2012). "Data Analysis Using Item Response Theory Methodology : An Introduction to Selected Programs and Applications." *Psychology Faculty Scholarship 20*.

Wisner, J. D., and Corney, W. J. (2001). "Comparing practices for capturing bank customer feedback." *Benchmarking: An International Journal 8*(3), 240–250.