Principles and Techniques for Creating and Validating Computer-Adaptive Surveys (CAS)

By

Sahar Sabbaghan

A thesis submitted in fulfilment of the
requirements for the degree of

Doctor of Philosophy
in Information Systems,

The University of Auckland, 2018.

ABSTRACT

Diagnostic theories are fundamental to IS practice and are represented in trees. One tool for representing diagnostic theories is Computer-Adaptive Surveys. Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the responses of previous questions asked. Their principal advantage is they allow the survey developer to input a large number of potential causes.

Respondents then roll down through the causes to identify the one or few significant causes impacting a correlation. As an example, consider a scenario where a café owner wants to know not only what aspect of customer service he could improve on, but also how he can improve. With traditional survey techniques, he would be forced to give customers a very long survey to identify the salient issues for improvement. In long surveys, respondents will suffer from a fatigue effect, and not answer questions properly.

As CAS items have certain characteristics, such as being multi-dimensional and containing constructs with parent-child relationships, traditional methods for assessing construct and conclusion validity are not suitable. For this thesis, I have developed techniques and principles for creating and validating CAS which is applied to a CAS on café satisfaction. First, I created a variant q-sorting methodology for assessing the construct validity of the CAS tree. In that method, tree 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 tree. The trees are then successively transformed to test if they branch in the same way. Second, the inter-rater reliability of the

trees of raters need to be assessed. However, existing ways of measuring inter-rater reliability such as the use of Cronbach's Alpha or the traditional way of using Cohen's Kappa don't work for CAS. I developed a way to measure inter-rater reliability in CAS using 3 measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma. Third, I assessed the conclusion validity of a café satisfaction CAS using two studies. The first study compared a café satisfaction CAS with online customer reviews and the second study compared it to traditional survey of the same item bank.

ACKNOWLEDGEMENTS

I take great pleasure in expressing my gratitude to all those who contributed to the inspiration and the knowledge required to complete this piece of work. I would like to express my gratitude to the following people whose time and efforts have contributed to this thesis.

To my supervisor, Associate Professor Cecil Eng Huang Chua, for his continuous support of my PhD study and research. For his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research, especially with the writing process. I remain amazed that despite his busy schedule, he was able to go through each section of this thesis numerous times. He would always respond within a few hours and
meet me every week with comments and suggestions on almost every paragraph.

To my co-supervisor, Associate Professor Lesley Ann Gardner, for her insightful comments and encouragement, but also for encouraging me to widen my research from various perspectives. I very much appreciate her guidance and all the useful discussions.

To Professor Michael Myers, for his great and valuable advice. I really appreciate his willingness to meet me on short notice and answer my never-ending questions.

To Professor David Sundaram, for his wise words of wisdom. Although his schedule is normally quite hectic, he always had time for a chat. I really appreciate his support throughout my studies.

To Elena Tang, who has always cheered me on and gave me great advice.

To Koro Tawa, for helping me with the MS SQL server. He has always been so kind to help me out on numerous occasions, especially in times of the server crashing.

To my friends at University of Auckland, especially, Miriam, Tianyou, Smita, Mariaelena and Dana who have all motivated me to strive towards my goal. I am fortunate to have met such wonderful people.

To all my close friends outside the university, especially Shahper, Sonia, Susan, and Sahar who always gave me their moral support, kindness, and wisdom.

To the participants of my research, I would like to sincerely thank and acknowledge all those, too many to be mentioned individually here, who were patient enough to go to my website and go through the surveys.

Finally, to my family, especially my loving husband, Chris, whom I am grateful for always being there and for his invaluable love, patience, support, and encouragement throughout this experience. This journey would not have been possible if not for him, and I dedicate this milestone to him.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ORIGINAL ARTICLES

This thesis consists of an introduction and conclusion, and three journal papers. The original publications are listed below along with a description of the author's contribution.

Journal Article I

Sabbaghan, S., Gardner, L., & Chua, C. E. H. (2018). Statistical Measurement of Tree's Similarity. *Submitted to Decision Support Systems.*

Abstract: Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. One way of performing conclusion validity on CAS trees is to employ independent raters to construct such trees and compare them. However, good measures of similarity to compare CAS trees have not been identified. This paper presents an analysis of the suitability of various measures of association to determine the similarity of two CAS trees using bootstrap simulations. We find that 3 measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954) each behave differently depending on what is inconsistent between two trees thus providing both measures for assessing alignment between two trees developed by independent raters as well as identifying the causes of the differences.

An abridged version of this article was published at ECIS 2017 in Guimaraes, Portugal.

Sabbaghan, S., Gardner, L., & Chua, C. E. H. (2017). A Threshold for a Q-Sorting Methodology for Computer-Adaptive Surveys. In *ECIS 2017 Proceedings*.

As first author, Sahar Sabbaghan has taken the lead in writing this article with editing done by the co-authors.

Journal Article II

Sabbaghan, S., Gardner, L., & Chua, C. E. H. (2018). A Variant Q-sorting Methodology for Building Trees of Psychometric Constructs. *In the process of submitting to MIS Quarterly.*

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 evaluating the validity of a CAS tree. This paper describes the process of using a variant of q-sorting to validate a CAS tree. In this methodology, trees 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 tree. The trees are then successively transformed to incrementally test if they branch in the same way. The results help researchers not only identify quality items for use in a CAS tree, but also facilitate diagnoses of problems with those items.

An abridged version of this article was published at ECIS 2016 in Istanbul, Turkey:

Sabbaghan, S., Gardner, L., & Chua, C. E. H. (2016). A Q-Sorting Methodology for Computer-Adaptive Surveys. *ECIS 2016 Proceedings*.

As first author, Sahar Sabbaghan has taken the lead in writing this article with editing done by the co-authors.

Journal Article III

Sabbaghan, S., Gardner, L., & Chua, C. E. H. (2018). Quantitative Instrumentation for Answering "Why" Questions: An Empirical Test. *Submitted to Marketing Science.*

Abstract: Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. Their principal advantage is they allow the survey developer to input a large number of potential causes. Respondents then roll down through the causes to identify the one or few significant causes impacting a correlation. This paper attempts to demonstrate that CAS are superior instruments for diagnosing root cause. To do this, we compare the conclusion validity of a CAS about café customer satisfaction against two other techniques, specifically (1) a psychometric survey, and (2) a content analysis of online customer reviews. We find that CAS has several advantages over the psychometric survey, as CAS has a higher response rate, requires fewer items for respondents to answer, which reduces fatigue effect, has better item discrimination in that respondents tend to provide more extreme scores in CAS, and has a higher agreement among respondents for each item. In addition, our results suggest CAS is a more robust approach to assessing café satisfaction than online reviews.

An abridged version of this article was published at PACIS 2017 in Langkawi, Malaysia and at ECIS 2018 in Portsmouth, United Kingdom.

Sabbaghan, S., Gardner, L., & Chua, C. E. H. (2016). A Test of a Computer-Adaptive Customer Satisfaction Survey using Online Reviews. *ECIS 2018 Proceedings*.

Sabbaghan, S., Gardner, L. A., & Chua, C. (2017). Computer-Adaptive Surveys (CAS) as a Means of Answering Questions of Why. *In PACIS Proceedings*, 1–15.

As first author, Sahar Sabbaghan has taken the lead in writing these articles with editing done by the co-authors.

1 Introduction

Diagnostic theories are fundamental to IS practice as in many cases we would want to know why the problem occurs to not only solve the problem but also for effective problem prevention (Rooney & Van den Heuvel, 2004). There have been numerous studies which have discussed the importance of finding the root cause of a phenomenon. One example is that when an IT system fails, it can fail for a myriad of reasons, such as disruptions caused by bugs (Grottke et al., 2016) or lack of efficient and effective designs (Dalal & Chhilar, 2013) or when change management process is not followed (Markus, 2004) or lack of management and organization support such as poor employee training (Yoon, Guimaraes, & O'Neal, 1995). Another example is that there can be many reasons why users do not adopt a technology (Bagozzi, 2007; Legris, Ingham, & Collerette, 2003). Developers of the technology have to trace through the factors to determine which impedes adoption. Finally, some IS systems, particularly expert systems, are inherently diagnostic. For instance, Mycin (Shortliffe, 2012) and other expert systems navigate a decision tree to identify the root cause of a problem. As an example, Hopp, Iravani, and Shou (2007) used a diagnostic tree for evaluating and improving production line performance. In many cases, we want to know not just what the problem is but also why such problems exist and identify the root cause.

In contrast, variance theories incorporate independent variables that statistically explain variations in the dependent variables (Webster & Watson, 2002). As an example, variance theories can be useful for identifying factors that increase the use of technology in workplace (Viswanath & Hillol, 2008). The relationship between variables or sets of variables are assessed by methods as factor analysis, ANOVA, and regression (Hair et al., 1998). Diagnostic theories are represented in trees whereas variance theories are represented as flat structures. In terms of survey design, variance theories are normally represented in psychometric surveys, while Computer-Adaptive Surveys are used to represent diagnostic theories.

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where large constructs with several interdependent dimensions are "unpacked" and mapped into a tree. The items thus have a parent-child relationship. The tree is then transformed into a survey where items asked of respondents depend on the responses of the previous questions asked. CAS can be useful in situations where there are two or more constructs which are correlated, and one desires to understand the reason why those constructs are correlated. CAS's main advantage is that it allows the survey developer to include a large number of questionnaire items. The only items the respondent answers are the ones most salient to the issue being addressed. The goal of CAS is to identify the root cause of the problem which is perceived by the respondent as most relevant to them. (e.g., which things did you like the least or most).

In CAS, large constructs are unpacked to explore and identify the different dimensions. Each of these dimensions is represented in an item, where each item in CAS reflects a potential root cause. These items are then mapped in a CAS tree. Items concerning higher level concepts link to items with greater detail. The CAS tree is then transformed into a survey to test which of these items are most relevant to the context of the study. For instance, consider an example in café satisfaction, where all the possible reasons why one would be unhappy with a café is explored and identified and mapped into a CAS tree. The only items the respondent answers are the ones most salient to the issue being addressed. Hence, CAS identifies the root cause of the problem which is perceived by the respondent as most relevant to them. (e.g., which things did you like the least or most). The results of CAS not only indicate which large top-level construct is an issue but precisely which dimension. Hence, If the respondent indicates that food was the issue the respondent was most dissatisfied with, then the CAS would proceed to administer questions on quality of food, while omitting further questions on price. The process repeats until it has drilled down to the specific issue with the food and CAS stops. Moreover, CAS can be efficient, since only questions most relevant to a respondent's concerns are asked,

hence only a fraction of the items are answered. Whereas in long questionnaires, the respondent is likely to encounter fatigue and quit before providing critical information (Galesic & Bosnjak, 2009; Groves, 2006b; Groves et al., 2004; Heerwegh & Loosveldt, 2006; S. R. Porter, Whitcomb, & Weitzer, 2004). CAS offers quantitative positivists a way to approach a domain traditionally held by qualitative researchers, i.e., the ability to answer questions of why.

1.1 Assessing the Validity of CAS

A survey instrument needs to be valid and reliable (Groves et al., 2004; Hayes, 1992). However, CAS requires different validation techniques as it differs from psychometric surveys. As an example, a CAS may have hundreds of items where the items are mapped into a tree, such as items higher up the tree are formatively defined by items lower down the tree. The traditional q-sort (Block, 1961) suffers limitations for managing CAS items, notably an inability to manage trees, as these methods assume a "flat" set of items. Another example, is that CAS uses an adaptive version of branching for respondents to move from one set of items to another set (Norman & Pleskac, 2002). Therefore, a range of possible routes is possible. As a result, the aggregated results produce both the frequency of the chosen items and their mean scores. For CAS to have credible results would mean that the "correct construct(s)" have been identified. Thus, unlike psychometric survey where the relationship between variables or sets of variables are assessed by using a variety of tools, such as structural equation modeling (SEM) techniques, partial least squares (PLS), and factor analysis (Hair, Anderson, Tatham, & Black, 1998), CAS requires different techniques. Unfortunately, methods for validating CAS are under researched- to the point where they are non-existent.

Validity can be defined as how well a tool measures what it sets out to measure (Litwin, 1995). There are clearly other forms of validity, such as construct, internal, and external validity

(Shadish, Cook, & Campbell, 2002). This study solely focuses on assessing the validity of the CAS tree structure and results of CAS.

The validation is performed in two stages, one the stage of pre-test which is the instrument development and two the post-test which is after data collection to ensure the credibility of the results.

In the first stage, the aim is to validate the CAS tree so that the correct constructs are mapped in the tree. Our technique not only assesses the construct validity of the tree, but also diagnoses which problem (mapping of items to item) causes problems with inter-rater agreement.

In the second stage the credibility of the results of CAS need to be assessed. CAS typically identifies one or a few narrowly defined constructs that respondents as a whole have the greatest or least affiliation to. Hence for CAS to have credible results would mean that the "correct construct(s)" have been identified. Thus, conclusion validity for CAS would mean that the results need to be checked and assessed against an external criterion which is a direct and independent measure of what the CAS is designed to measure.

1.2 Motivation

Despite that diagnostic theories are core to IS practice, little attention has been paid in IS research to diagnostic theories. Perhaps this is because diagnostic theories are not explanatory theories like variance or process theories, but instead are prescriptive theories- they directly inform decision making. IS traditionally has focused on explanation rather than design and decision making (Markus & Robey 1984, 1988). Nevertheless, like all theory, diagnostic theories need to be validated. Computer-Adaptive Surveys is a new tool for representing diagnostic theories. This study aims to develop techniques and principles for creating and validating a CAS.

As, demonstrated in section 1.1, there are many different kinds of validity. This study focuses on one, validating the CAS tree structure as this is the first step of creating a CAS as nothing else can be done until this is solved and two, assessing the results of a CAS against an external criterion(s), as the ability and efficacy of CAS (i.e., conclusion validity) for actually unpacking variables/constructs to answer questions of why remains unknown. Thus, this thesis aims to do the following:

1. Develop a method to compare two CAS trees

2. Create a CAS tree by using a variant q-sort technique which is designed to assess the validity of CAS trees

3. Assess the ability and efficacy of CAS

1.3 Glossary of Key Concepts

This study contains certain terms which will be used consistently throughout the paper. The following provides a glossary of what each term means. Further explanation will be given in the next chapters.

Ancestor: an ancestor is the *n-th* degree parent of a descendant node where $n>0$. As an example, the ancestor of 4^{th} degree is located 4 degrees above its descendant node. A first-degree ancestor of a node is also called the parent node.

Construct: is defined as a "concept, model, or schematic idea" (Shadish et al. 2001, p. 506). Formative constructs are defined by the dimensions or measures that form them (Petter, Straub, & Rai, 2007).

Concept: is defined as a formally and logically developed idea of a phenomena.

Descendant: is the *n-th* degree child of an ancestor node. As an example, the descendant of a *3rd* degree of the node is located 3 degrees below its ancestor node. A first-degree descendant of a node is also called the child node.

Degree: given two nodes *a* and *b* of level *m* and *n* such that a is the ancestor of *b* or *b* is the ancestor of *a*. The degree of the pair $d(a,b) = |m-n|$.

Diagonal movement: is a direction movement *M(a,b*) that is both a hierarchy and level movement. Diagonal movements suggest two raters thought of a construct in very different ways, as they disagree on both the level of the mapping and their direct relative constructs.

Diagonal swap: is a swap where *a* and *b* in *T* are located in different levels and are not relatives

Dimension: A construct C has more than one dimension if there exist multiple measures of the construct M1, M2, .. Mn such that M1, M2, .. Mn are not correlated, and Z=sum(M). In a CAS tree, higher level constructs represent concepts that are more general and have more dimensions, whereas, lower level constructs represent concepts that are more in depth and have fewer dimensions.

Direction of movement: given two nodes *a* and *b* in tree *T*, *a* can move in three possible directions to become a child of *b* in *T'*, (1) within relatives (up or down), (2) within its level (left or right), or (3) both within relatives and levels. Movements can occur with any kind of node (with or without descendant).

Hierarchy movement: is a direction movement *M(a,b)* where *a* and *b* are nodes in *T* and *a* is a relative of *b*. In *T' a* becomes a first-degree descendant of *b*. If *b* is descendant of *a* in *T*, then a hierarchy movement type 3 is not possible, because effectively, nothing happens to *b*. A hierarchy movement is effectively a movement up or down the tree.

Hierarchy swap: is where *a* is a relative of *b* in *T*.

Level: is the distance of a node from the root. A node is on the *n+1* level of its parent node. As an example, a node located on the 3rd level is 3 levels below the root node and its parent is on the 2nd level. Levels closer to the root are considered as higher levels and levels further from the root are considered as lower levels.

Level movement: is a direction movement *M(a,b) a* is in the same level as a child of *b*, is defined as: *T*' such that *a* is the child of *b*. We care about level movements, because these suggest that while raters agree on the level of the construct, they disagree on the "family" of constructs the question relates to.

Level swap: is a swap where *a* and *b* in *T* are located in the same level and if both *a* and *b* do not have any descendants, they must not share a first-degree ancestor (direct parent) as tree T' will be the same as T.

Measures: also known as **indicators** or **items**, are observable, quantifiable scores obtained through self-report, interview, observation, or other empirical means (Edwards and Bagozzi 2000). Hence, to simplify things I only use the term items.

Modification: given two trees, one of the differences between the two trees.

Movement: Given a tree *T* with nodes labelled from *1* to *n*. A movement *M(a,b)* where *a* and *b* are nodes in *T* such that *0<=a<=n, 0<=b<=n, a<>b* and *b* has descendants and *a* is not a first-degree descendant of *b*, is defined as: *T*' such that *a* is the child of *b* , i.e., *a* is a first-degree descendant of *b*.

There are three types of movements:

- **Type 1 movement:** is a movement such that *a* in *T* is a childless node.
- **Type 2 movement:** is a movement such that in *T*, *a* is a parent node. In *T'*, all descendants of *a* in *T* become descendants of *a's* parent.

- **Type 3 movement with child(ren):** is a movement such that in *T*, *a* is a parent node. In *T'* all descendants of *a* in *T* have the same parents.

Non-Relative: is any node whose only ancestor to another node is the root.

Root: is the node with no parent. The root of the tree is on level 0.

Relative: Given two nodes, *a* and *b*, either a and *b* share an ancestor which is not the root, or *a* is an ancestor of *b*, or *b* is an ancestor of *a*.

Swap: a combination of two or more movements, which we treat as one. A swap is *S(a,b)* in *T*, where in *T', b* becomes the child of *a's* parent and takes *a's* children as descendants (if any), while *a* becomes the child of *b'*s parent and takes *b*'s children as descendants (if any). We consider swap distinctive from movements because this reflects a single cognitive difference between two raters rather than two or more cognitive differences. Similar to direction movements, there are three types of swaps, which are hierarchy, level, and diagonal.

Top level: are first-degree descendants of the root. The top level of the tree has a level of 1.

Validity: can be defined as how well a tool measures what it sets out to measure (Litwin, 1995). There are several forms of validity, such as construct, internal, and external validity (Shadish, Cook, & Campbell, 2002).

1.4 Thesis Structure

In this research, I have developed principles for creating and assessing Computer-Adaptive Surveys (CAS). Computer Adaptive Surveys are a new form of survey which I have developed. This thesis aims to develop principles and techniques for developing computer-adaptive surveys (CAS) for measuring latent psychometric properties. There are clearly other forms of validity, such as internal, and external validity (Shadish, Cook, & Campbell, 2002). In this thesis, I specifically intend to develop principles for creating and assessing the validity of CAS

trees and assessing the accuracy and credibility of CAS results. I employed café satisfaction as a proxy context to demonstrate the usefulness of the principles and techniques.

I chose café satisfaction instead of a more salient IS topic (e.g., TAM) for several reasons. One, because of the availability of existing, relevant instruments, as if I were to develop our own TAM/UTUAT instrument, the potential quality of the questionnaire items would be a confound. By using existing items employed in traditional surveys, this confound is eliminated. Two, there are numerous studies which have provided theoretical frameworks (Liang & Zhang, 2009; Shanka & Taylor, 2005). Three, there are many reviews and blogs on cafes which makes assessing the credibility of the results more feasible. Finally, "cafes" is a topic that anyone can relate and understand which makes the task of finding independent and blind raters more practical.

Chapter 2 presents an analysis of the suitability of various measures of association to determine the similarity of two CAS trees using bootstrap simulations. In Chapter 3, CAS trees employs independent raters to construct such trees and compare them. However, good measures of similarity to compare CAS trees have not been identified. I find that 3 measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954) each behave differently depending on what is inconsistent between two trees thus providing both measures for assessing alignment between two trees developed by independent raters as well as identifying the causes of the differences.

Next, Chapter 3 describes the process of using a variant of q-sorting to validate a CAS tree. In this methodology, trees 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 tree. The trees are then successively transformed to incrementally test if they

branch in the same way. The results help researchers not only identify quality items for use in a CAS tree, but also facilitate diagnoses of problems with those items.

Chapter 4 attempts to demonstrate that CAS are superior instruments for diagnosing root cause. To do this, I compare the conclusion validity of a CAS about café customer satisfaction against two other techniques, specifically (1) a psychometric survey, and (2) a content analysis of online customer reviews. I find that CAS has several advantages over the psychometric survey, as CAS has a higher response rate, requires fewer items for respondents to answer, which reduces fatigue effect, has better item discrimination in that respondents tend to provide more extreme scores in CAS, and has a higher agreement among respondents for each item. In addition, my results suggest CAS is a more robust approach to assessing café satisfaction than online reviews.

Finally, I conclude this thesis by first providing the summary of the contributions of each chapter. In addition, some opportunities for future research is presented.

References

Bagozzi, R. P. (2007). The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift. Journal of the Association for Information Systems, 8(4), 244–254.

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.

Dalal, S., & Chhillar, R. S. (2013). Empirical study of root cause analysis of software failure. *ACM SIGSOFT Software Engineering Notes*, *38*(4), 1.

Dennehy, D., & Conboy, K. (2017). Going with the flow: An activity theory analysis of flow techniques in software development. *Journal of Systems and Software*, 133, 160–173.

Galesic, M., & 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.

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classifications. *Journal of the American Statistical Association*, *58*(302).

Grottke, M., Kim, D. S., Mansharamani, R., Nambiar, M., Natella, R., & Trivedi, K. S. (2016). Recovery from failures caused by mandelbugs in IT systems. *IEEE Transactions on Reliability*, *65*(1), 70–87.

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., & 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).
Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis*. (J. F. Hair, R. E. Anderson, R. L. Tatham, & W. C. Black, Eds.), *International Journal of Pharmaceutics* (Vol. 1). Prentice Hall.
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.
Hopp, W. J., Iravani, S. M. R., & Shou, B. (2007). A Diagnostic Tree for Improving Production Line Performance. *Production and Operations Management*, 16(1), 77–92.
Liang, X., & Zhang, S. (2009). Investigation of customer satisfaction in student food service. *International Journal of Quality and Service Sciences*, *1*(1), 113–124.
Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. Information & Management, 40(3), 191–204.
Litwin, M. S. (1995). *How to Measure Survey Reliability and Validity*. Thousand Oaks ; London: Sage.
Markus, M. L., & Robey, D. (1988). Information technology and organizational change: causal structure in theory and research. *Management Science*, *34*(5), 583-598.
Norman, K. L., & Pleskac, T. (2002). Conditional branching in computerized self-administered questionnaires on the World Wide Web. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*, 1241–1245.
Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, *20*(1), 61–71. Petter, S., Straub, D., & Rai, A. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, *31*(4), 623–656.
Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, *2004*(121), 63–73.
Robey, D., & Markus, M. L. (1984). Rituals in information system design. *MIS quarterly*, 5-15.
Rooney, J. J., & Van den Heuvel, L. N. (2004). Root cause analysis for beginners. *Quality Progress*, (July), 45–53.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental for Generalized Designs Causal Inference*. *Wadsworth Cengage learning*. Houghton Mifflin Company.
Shanka, T., & Taylor, R. (2005). Assessment of university campus café service: The students' perceptions. *Asia Pacific Journal of Tourism Research*, *10*(March), 329–340.
Shortliffe, E. (2012). Computer-based medical consultations: MYCIN (Vol. 2). Elsevier.
Viswanath, V., & Hillol, B. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315.
Yoon, Y., Guimaraes, T., & O'Neal, Q. (1995). Exploring the factors associated with expert systems success. *MIS Quarterly*, 19(1), 83–106.
Webster, J., & Watson, R. T. (2002). Analyzing the Past To Prepare for the Future : Writing a Review. *MIS Quarterly*, *26*(2).

2 Statistical Measurement of Trees' Similarity

Abstract

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. One way of performing construct validity on CAS trees is to employ independent raters to construct such trees and compare them. However, good measures of similarity to compare CAS trees have not been identified. This paper presents an analysis of the suitability of various measures of association to determine the similarity of two CAS trees using bootstrap simulations. We find that 3 measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954) each behave differently depending on what is inconsistent between two trees thus providing both measures for assessing alignment between two trees developed by independent raters as well as identifying the causes of the differences.

Keywords: Computer-adaptive survey, construct validity, q-sorting, threshold

2.1 Introduction

A tree is defined by a finite, connected, directed graph with no loops, only one node with outdegree 0, namely the root, and all other nodes with outdegree 1 (Geoffrion, 1989). Trees are among the most common and well-studied structures and the problem of comparing trees occurs in diverse areas such as text recognition (Heilman & Smith, 2010), network data (Clauset, Moore, & Newman, 2008), security evaluations (Yamakoshi, Tanaka, & Iwasaki, 2016) and databases (Wotring & Ripley, 2005). The problem of comparing trees occurs in diverse areas such as text recognition (Heilman & Smith, 2010), network data (Clauset, Moore, & Newman, 2008), security evaluations (Yamakoshi, Tanaka, & Iwasaki, 2016) and databases (Wotring & Ripley, 2005). There have been several approaches for comparing trees. One popular method is to measure the number of changes to convert one tree into another, i.e., tree edit distances (Tai, 1979). Example algorithms include Klein (1998) which focused on two

ordered trees, Chen (2001) which proposed a new algorithm for computing the minimum edit cost of transforming one ordered tree to another one and Jiang et al.(1995) which measures the similarity between two ordered and unordered trees.

However, these methods are not suitable for comparing all types of trees. As an example, consider using independent raters to construct-validate a Computer-Adaptive Survey (CAS) tree. A CAS can have hundreds of constructs, which independent raters arrange in a tree with constructs concerning higher level concepts linking to constructs with greater precision. As an example, consider a café satisfaction CAS tree. When respondents indicate lack of satisfaction with a construct like "food quality," the CAS asks them to evaluate subconstructs like "taste of food" and "portion sizes of the food."

To use independent raters, a way must exist to measure if two CAS trees are "similar enough." Methods such as edit distance have limitations for such problems, because they are sample size dependent. An edit distance of 20 is very bad when comparing two trees with 40 nodes each but is not so bad if the two trees have over 1000 nodes. A more robust measure of the similarity between two trees that can be employed for statistical comparison like Cohen's Kappa is therefore needed.

To address our problem, we performed a set of bootstrap simulations to measure how various statistics change as a hypothetical CAS tree deviates from a "true" version. We find that the 3 measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954) together provide information useful for assessing construct validity in CAS. Each of these three statistics behave differently depending on what is inconsistent between two CAS trees thus providing both metrics for assessing alignment between two CAS trees developed by independent raters as well as identifying the causes of the differences

The paper is constructed in the following manner. First, we introduce CAS and present the properties of CAS trees and limitations of previous work. Then, we attempt to address those problems by providing a process for developing good thresholds for the construct validity of CAS and diagnosing their differences. Finally, we conclude the paper with a discussion of diagnosing inter-rater reliability and conclusion.

2.2 Introduction to Computer-Adaptive Surveys (CAS)

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the responses of the previous questions asked. Computer-Adaptive Surveys (CAS) are most useful for the situation where there are two or more constructs which are correlated, and one desires to understand the reason why those constructs are correlated. Its principal advantage is that it allows the survey developer to include a large number of questions. The only questions the respondent answers are the ones most salient to the issue being addressed- in our running example, the things about the café the respondent is least satisfied with. In contrast, if the same number of questions were asked on a traditional survey, the respondent is likely to encounter fatigue and quit before providing critical information (Galesic & Bosnjak, 2009; Groves, 2006; Groves et al., 2004; Heerwegh & Loosveldt, 2006; Porter, Whitcomb, & Weitzer, 2004). Methods such as Structural Equation Modeling (SEM) are used to examine the causal relationships and test the hypotheses between the observed and latent items in a research model (Hoyle, 1995). However, in many situations one may desire to know why such relationships exists. For instance, Bagozzi (2007) highlights that research on the technology acceptance model has demonstrated the relationship between perceived usefulness, ease of use, and intention to use, but cannot articulate why this relationship holds. Most IS survey research has been useful for developing causal relationships between constructs, but has been poor at "unpacking" constructs to develop a rich understanding of how things work. A CAS tree first provides a theoretical framework for all the possible reasons why

a relationship exists between two variables and then that framework is tested by transforming the tree into a CAS. The CAS identifies which of those reasons is most relevant to the context being tested. Hence, CAS enables IS researchers to begin quantitatively exploring questions of why, thereby furthering theory (Sutton and Staw 1995).

In CAS, respondents perform a depth-first traversal of the tree, where each stage of the traversal involves the respondent rating all items in the stage. Respondents then receive only child constructs associated with the lowest or highest rated constructs. Each respondent could traverse the CAS tree in a different way. To illustrate, see Figure 2.1. If food is the area the customer is least satisfied with, CAS retrieves questions about food (i.e., preparation, portion, menu choice). If the customer is least satisfied with menu choice, CAS then retrieves questions about the special needs, options, and availability of the food. CAS does not retrieve further questions on constructs the respondent rated satisfactorily. As the respondent continues to answer questions, CAS navigates deeper down the tree and questions roll down until the respondent hits one set of constructs with no children, for example, that there are insufficient vegetarian items on the menu. If most respondents agree there are insufficient vegetarian items on the menu, this would indicate lack of vegetarian items is the root cause for many customers' dissatisfaction. Of course, not every respondent would navigate the tree the same way. Thus, aggregating the results from respondents allows the researcher to observe the multiple major problems across respondents. In addition, it allows managers of commercial enterprises to quickly find key issues to address.

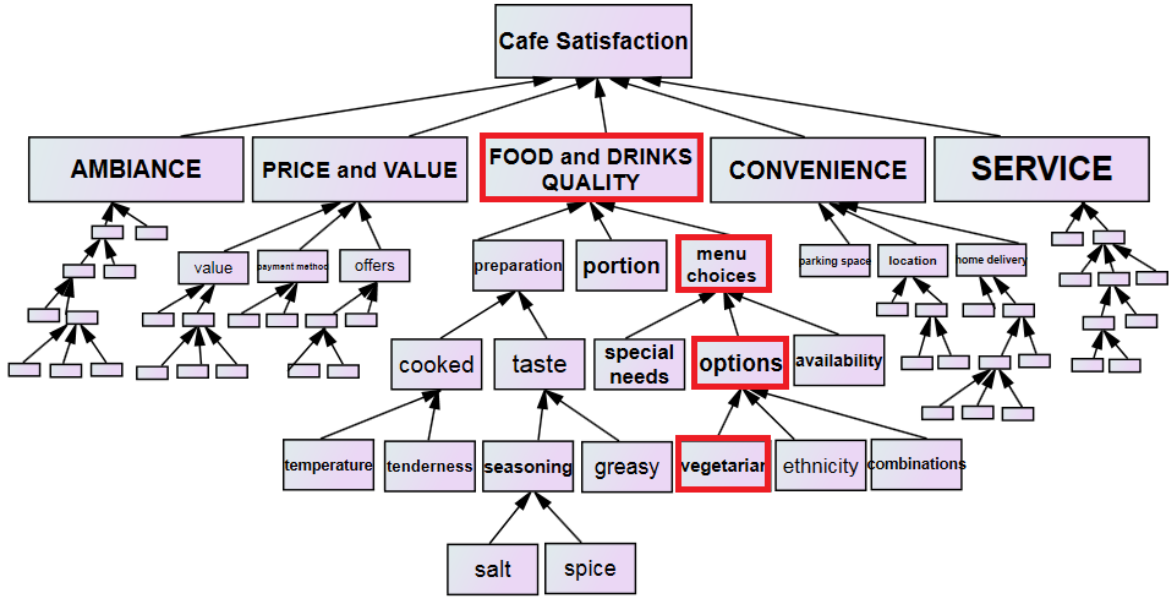


Figure 2.1. How CAS works for café satisfaction

Comparison of CAS Trees

The trees being compared in CAS have certain properties. First, a CAS tree can have hundreds of constructs, where constructs concerning higher level concepts are mapped to constructs with greater precision. The constructs then have a parent-child relationship. Second, a proportion is not sufficient to represent the number of nodes, as this relies on the assumption that the error is linear, as a 20-node tree with 2 problematic nodes is not the same as a 200-node tree with 20 problematic nodes. As in CAS, constructs higher up the tree are more important than those lower in the tree. This is because constructs lower in the tree are subconstructs of those higher in the tree. This means that if there are any errors or disagreement in higher levels, then those errors will continue in propagate to lower levels, as if raters disagree in higher levels, they will most likely to disagree on the lower level constructs. Thus, errors higher in the tree cascade down.

CAS trees are constructed through a variant q-sorting methodology. In the methodology, trees 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 nodes in the hierarchy. The trees are then successively transformed to test if they branch in the same way.

However, the aforementioned q-sorting methodology has two key limitations. One, what are appropriate measures and "good enough" thresholds for demonstrating the similarity of two CAS trees remains unknown. Second, good measures for identifying problematic items in the tree are undiscovered. In a CAS tree, higher level constructs represent concepts that are more general and have more dimensions. Whereas, lower level constructs represent concepts that are more in depth and precise and have fewer dimensions.

As an example, if two raters disagree on the mapping of two constructs, one would want to know whether raters consider that the constructs belong to different parents, or whether the raters disagree on the precision of the construct in the hierarchy, and hence is the construct located on the correct level of the tree hierarchy. The term precision in this context means level of depth of a construct. In effect, measures akin to the modification indices of variance-based structural equation models (Gefen, Straub, & Boudreau, 2000; Landis, Beal, & Tesluk, 2000) need to be formulated.

The remainder of this section reviews the principal existing methods of measuring tree similarity, which are edit-distance and statistics-based. We demonstrate the limitations of both methods and identify elements that we apply to provide a foundation for creating a threshold for CAS trees.

2.2.1 Edit-Distance Based Techniques

Edit distance is a poor general comparator for CAS trees for several reasons. One is that existing algorithms do not take into account that nodes in the tree are not equally important. To illustrate, consider Figure 2.2. In the figure, Trees B and C are each inconsistent with Tree A in exactly one way. Tree B swaps nodes 2 and 5, while Tree C swaps nodes 2 and 3. The typical edit distance algorithm treats both inconsistencies equally. However, Tree B suggests

the rater considered item 5 as a parent to item 2, while the rater for tree B considered items 2 and 3 to be completely different constructs from the rater creating Tree A. The difference represented in Tree C is more serious than Tree B, as the rater (1) considered the items to effectively be two different constructs (as opposed to different levels in the same construct), and (2) the issue occurred at a relatively high level in the tree, suggesting there are further problems lower in the tree that were undiscovered. In comparing CAS trees, a technique that identifies such differences is necessary.

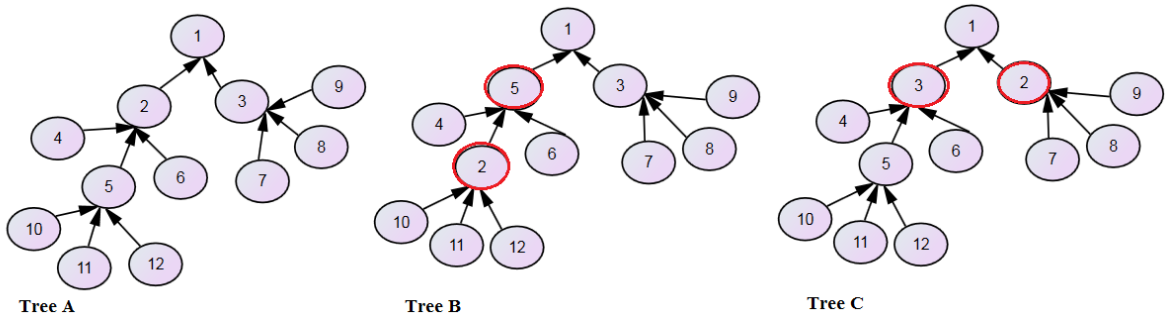


Figure 2.2. Tree hierarchy limitation example.

Also, edit distance measures are often sample size sensitive. Clearly if there are two trees, each having 50 items, where 10 changes are required to transform one into the other, this is different from having two trees, each having 500 items where only 10 such changes are required. In statistical thinking, we want to compare the statistic to some probability distribution to standardize results according to "node sample size." We then calculate confidence intervals or *p*-values of significance, where the threshold (typically 0.05 or 0.01) is sample size independent. The tree edit distance literature has no equivalent analogue.

Finally, as a corollary to the above two points, we would like our measures of tree similarity to systematically identify where the differences are between trees. Edit distance algorithms do this for individual nodes- they identify that to transform one tree into another, these are the nodes that must be changed and how. What they do not do is, for example, to tell us that most

of the errors are occurring in the top of the tree (very bad) or at the bottom of the tree (not so serious). Or to tell us that most of the errors are occurring in the children of construct A.

2.2.2 Statistics Based Techniques

Existing statistics-based methods are not suitable for several reasons. One is that existing measures and statistics are employed for generally "flat" question structures, and not the hierarchical structure of trees. For example, the traditional factor analytic concepts of convergent and divergent validity are assessed with correlations (Hair et al., 1998). However, in CAS trees, questions have a parent-child relationship. If the constructs behave correctly, the parent will correlate highly with at least one of the children but is unlikely to correlate with all. In formative constructs, not all child constructs may have a high correlation with their parent construct. Due to the direction of causality with formative models, high correlation between the items is not expected, required or a cause for concern (Diamantopoulos, Riefler, & Roth, 2008). For example, if a respondent answers that she is dissatisfied with the food quality, then the respondent might be unhappy about the way the food was prepared but be satisfied with the portion size. Factor loadings do not take this into account.

The inferential statistics tradition in several academic disciplines such as information systems is to employ thresholds to evaluate whether two things are the same or different (Boudreau, Gefen, & Straub, 2001). For example, we regularly consider a p-value under 0.05 to be "good enough." In cases where these thresholds are unknown, research is done to identify them. As an example, Hu and Bentler (1999) examine the adequacy of the "rules of thumb" of conventional cutoff criteria and propose several new alternatives for various fit indexes in structural equation models. However, so far, such techniques have not been applied to trees.

In this study, we apply traditional statistical measures in a new way to measure tree similarity. Essentially, we map the tree into a contingency table and employ traditional contingency table statistics to evaluate similarity. The three measures used are Goodman and Kruskal's Lambda,

Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954). Goodman and Kruskal (1954, p.749) interpret Lambda as "how much more probable it is to get like than unlike orders in the two classifications, when two individuals are chosen at random from the population." We chose Lambda (λ) because it has a meaning akin to r in a regression (Anderson & Gerbing, 1988), i.e., Lambda is the measure of strength of association in a contingency table (Everitt, 1992; Goodman & Kruskal, 1963). We chose Kappa, because it is widely used as a measure of association for contingency tables (Hambleton & Zaal, 2013; Rudick, Yam, & Simms, 2013; Sengupta & Te'eni, 1993; You, Xia, Liu, & Liu, 2012). According to Cohen (1968), Kappa is the observed proportion of agreement between the assigners after chance agreement is removed from consideration. In addition, Landis and Koch (1977) proposed the English-language meanings of Kappa thresholds featured in Table 2.1. We chose Gamma (γ) because it is explicitly designed for data with ordinal values, and hierarchies are ordered data structures. Goodman and Kruskal interpret Gamma (γ) as "how much more probable it is to get like than unlike orders in the two classifications, when two individuals are chosen at random from the population" (Davis, 1967; Göktaş & İşçi, 2011; Goodman & Kruskal, 1954). The value of the Gamma coefficient ranges from -1 to + 1 where the latter value indicates perfect agreement between the two classifications (Baker, 1974).

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Table 2.1. Kappa Interpretation.

2.3 Foundation for Threshold Building

To build suitable thresholds for comparing and assessing CAS trees, we first generate a hypothetical "perfect" tree. We then make a copy of the tree and systematically change the tree and measure the statistic. We make a second change on the tree and measure the statistic again.

We repeat the process many times. By doing this, we get a good appreciation for how statistics vary as two trees diverge. We then do the same with the other "perfect" trees of various sizes.

There is one constraint on the modifications, which is each parent cannot have just one child construct- with only one child construct, there is no "branching" and hence no choice. Nodes in CAS denote a respondent making a choice as to which construct is more or less favoured. In the below, we formally define certain terms we will employ in the remainder of the paper.

Level: is the distance of a node from the root. A node is on the *n+1* level of its parent node. As an example, a node located on the 3rd level is 3 levels below the root node and its parent is on the 2nd level. Levels closer to the root are considered as higher levels and levels further from the root are considered as lower levels.

Root: is the node with no parent. The root of the tree is on level 0.

Degree: given two nodes *a* and *b* of level *m* and *n* such that a is the ancestor of *b* or *b* is the ancestor of *a*. The degree of the pair $d(a,b) = |m-n|$.

Descendant: is the *n-th* degree child of an ancestor node. As an example, the descendant of a *3rd* degree of the node is located 3 degrees below its ancestor node. A first-degree descendant of a node is also called the child node.

Top level: are first-degree descendants of the root. The top level of the tree has a level of 1.

Ancestor: an ancestor is the *n-th* degree parent of a descendant node where $n>0$. As an example, the ancestor of 4^{th} degree is located 4 degrees above its descendant node. A first-degree ancestor of a node is also called the parent node.

Relative: Given two nodes, *a* and *b*, either a and *b* share an ancestor which is not the root, or *a* is an ancestor of *b*, or *b* is an ancestor of *a*.

Non-Relative: is any node whose only ancestor to another node is the root.

Modification: given two trees, one of the differences between the two trees.

Movement: Given a tree *T* with nodes labelled from *1* to *n*. A movement *M(a,b)* where *a* and *b* are nodes in *T* such that *0<=a<=n, 0<=b<=n, a<>b* and *b* has descendants and *a* is not a first-degree descendant of *b*, is defined as: *T*' such that *a* is the child of *b* , i.e., *a* is a first-degree descendant of *b*.

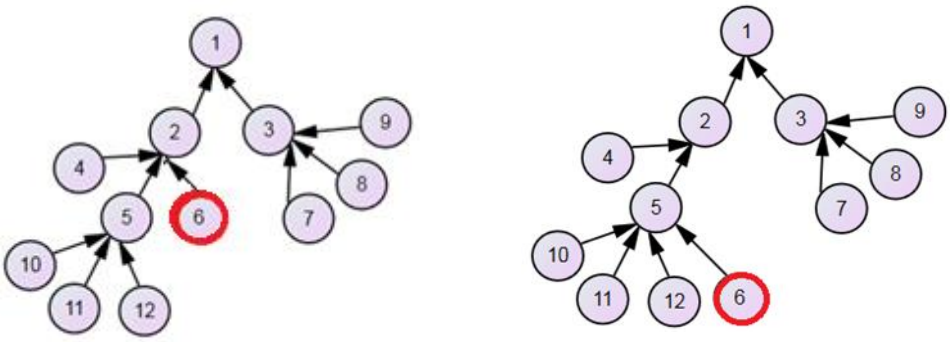
There are three types of movements:

- **Type 1 movement:** is a movement such that *a* in *T* is a childless node.
- **Type 2 movement:** is a movement such that in *T*, *a* is a parent node. In *T'*, all descendants of *a* in *T* become descendants of *a's* parent.
- **Type 3 movement with child(ren):** is a movement such that in *T*, *a* is a parent node. In *T'* all descendants of *a* in *T* have the same parents.

Direction of movement: given two nodes *a* and *b* in tree *T*, *a* can move in three possible directions to become a child of *b* in *T'*, (1) within relatives (up or down), (2) within its level (left or right), or (3) both within relatives and levels. Movements can occur with any kind of node (with or without descendant).

- **Hierarchy movement**: is a direction movement *M(a,b)* where *a* and *b* are nodes in *T* and *a* is a relative of *b*. In *T' a* becomes a first-degree descendant of *b*. If *b* is descendant of *a* in *T*, then a hierarchy movement type 3 is not possible, because effectively, nothing happens to *b*. A hierarchy movement is effectively a movement up or down the tree. We care about hierarchy movements, because these suggest a certain type of error. In a CAS tree, the top-level constructs are unpacked and their descendants are mapped. This type of error indicates that raters disagree on the mapping of their direct relative constructs. As an example, consider Figure 2.3 which presents two trees from raters 1 and 2. The raters disagree on the parent of construct 6 as rater 1 has mapped construct

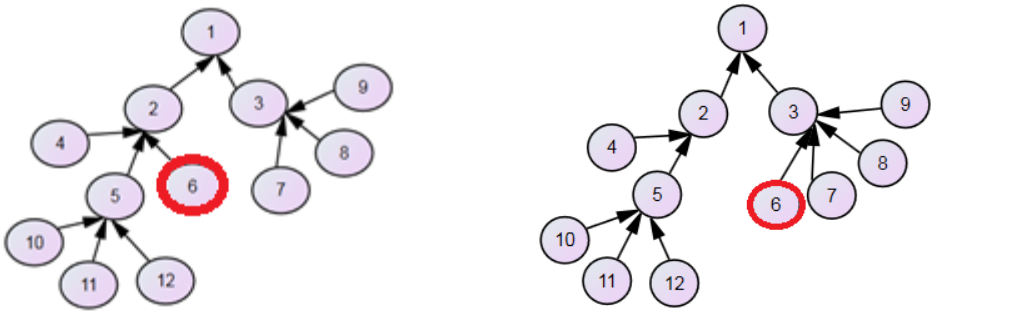
6 to construct 2, while rater 2 has mapped construct 6 to construct 5. In addition, rater 2 sees construct 6 in a lower level than rater 1, as rater 1 had mapped construct 6 to construct 2 which is in a higher level. In this example, as construct 6 (*a*) moved to become child of construct 5 (*b*), and does not have any descendants, we call this a type 1 hierarchy movement.



(a) rater 1	(b) rater 2

Figure 2.3. Example of type 1 hierarchy movement

- **Level movement**: is a direction movement *M(a,b) a* is in the same level as a child of *b*, is defined as: *T*' such that *a* is the child of *b*. We care about level movements, because these suggest that while raters agree on the level of the construct, they disagree on the "family" of constructs the question relates to. As an example, consider Figure 2.4(a) where in Tree 1, construct 6 is in the same level as constructs 7, 8, and 9. Figure 2.4(b) presents level movement type 1 for construct 6 (*a*) as it moved from construct 2 to construct 3 (*b*). The level of construct 6 has not changed, however raters disagree on the direct parent construct. This indicates that the raters are confused between constructs 2 and 3.



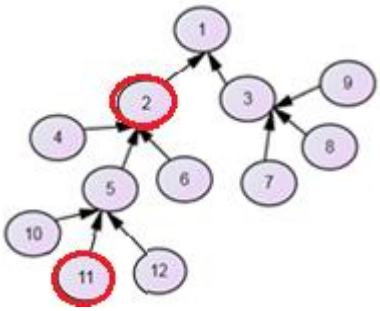
(a) rater 1	(b) rater 2

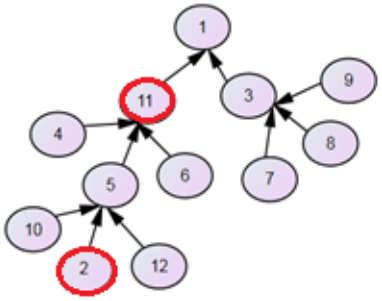
Figure 2.4. Example of level movement type 1.

- **Diagonal movement**: is a direction movement *M(a,b)* that is both a hierarchy and level movement. Diagonal movements suggest two raters thought of a construct in very different ways, as they disagree on both the level of the mapping and their direct relative constructs.

Swap: a combination of two or more movements, which we treat as one. A swap is *S(a,b)* in *T*, where in *T', b* becomes the child of *a's* parent and takes *a's* children as descendants (if any), while *a* becomes the child of *b*'s parent and takes *b*'s children as descendants (if any). We consider swap distinctive from movements because this reflects a single cognitive difference between two raters rather than two or more cognitive differences. Similar to direction movements, there are three types of swaps, which are hierarchy, level, and diagonal.

- **Hierarchy swap:** is where *a* is a relative of *b* in *T*. As an example, consider Figure 2.5, which presents a hierarchy swap between constructs 11 and 2. In *T'* construct 11 is closer to the root (construct 1), hence it becomes the ancestor of construct 2. This shows that raters disagree on the mapping of the direct relative constructs of constructs 2 and 11.

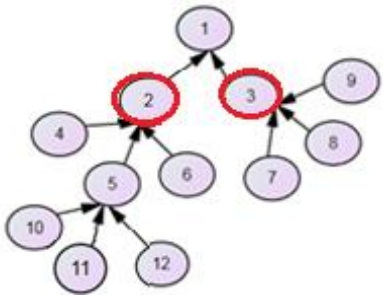


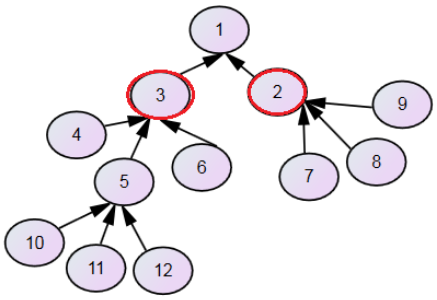


(a) rater 1	(b) rater 2

Figure 2.5. Example of hierarchy swap.

- **Level swap:** is a swap where *a* and *b* in *T* are located in the same level and if both *a* and *b* do not have any descendants, they must not share a first-degree ancestor (direct parent) as tree T' will be the same as T. As an example, in Figure 2.6, constructs 2 and 3 are on the same level and have been swapped. This shows while raters agree on the level of the constructs, they disagree on the mapping of the parent construct.





(a) rater 1	(b) rater 2

Figure 2.6. Example of level swap.

- **Diagonal swap:** is a swap where *a* and *b* in *T* are located in different levels and are not relatives. Diagonal swaps suggest the two raters are confused with two constructs in very different ways, as they disagree on both the level of the mapping and their direct relatives.

In addition, we want to perform analyses comparing the result of moving nodes at higher levels of the tree versus moving nodes at lower levels of the tree. Changing nodes at higher levels of

the tree should have a greater impact, because this suggests problems with more important constructs. As an example, consider Figure 2.7, which presents Tree A, B and C. In Tree A and B, raters disagree in the mapping of construct 3 and in Tree A and C the raters disagree on the mapping of construct 14. The disagreement between Tree A and B is more serious than the disagreement between Tree A and C.

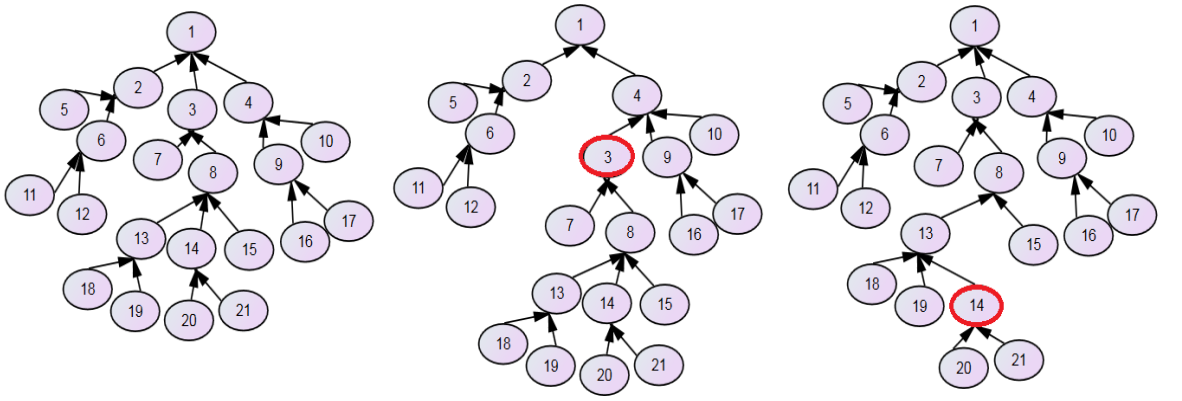


Figure 2.7. Example of levels in CAS trees.

To simulate these conditions, we perform analyses where we restrict the levels that movements and swaps occur at. In our analysis, for every perfect tree with levels *0..n*, we introduce a variable *x* where $2<x<n$. Using the perfect tree as a base, we perform a set of swaps and movements between levels *1* and *x*. We then use the perfect tree as a base again, and perform a second set of swaps and movements between levels *x* and *n*. We compare the difference in scores between these two. To distinguish the two, the swaps and movements performed between levels *1* and *x* are called movements and swaps on the "top" of the tree, and those between *x* and *n* as on the "bottom" of the tree.

Insertion and Deletion

Finally, in some cases, one rater may not choose to map all of the pre-determined constructs and the two trees could have different numbers of nodes. Hence, we assess the impact of an insertion/deletion of a node in a tree. As deletion is the reverse process of insertion of a node,

we only consider assessing the impact of insertions. We consider two types of insertion as there are only two ways to insert a node to a tree, (1) insertion in levels where there is an increase in the number of branches per node and (2) insertion in hierarchy where there is an increase in the number of levels.

2.3.1 Diagnosing Process and Threshold Building

In this study, for the diagnosing process, we systematically identified all the ways a construct can move in a tree hierarchy which has 1 to n- levels and 1 to m branches per node. We generate 12 (i.e., 3 x 4) perfect trees to test. Each tree has between 3-5 (i.e., 3 possibilities) branches per node and 3-6 levels (i.e., 4 possibilities) as tabulated in Table 2.2. We did not perform the simulation on trees with more than 200 nodes, because the computations required to simulate these trees become exponentially complex (Goldreich, 2011). Our analysis on smaller trees suggest that statistics are similar regardless of the size of the tree. In addition, we drop the 3-branch 3 level tree as the number of constructs is too small to run a simulation on for 100 rounds. Hence, 6 trees remain for the diagnosing process. These are identified in as the grey shaded cells Table 2.2.

BRANCHES	**LEVELS**			
	3	4	5	6
3	13	40	121	364
4	21	85	339	1363
5	30	156	781	3906

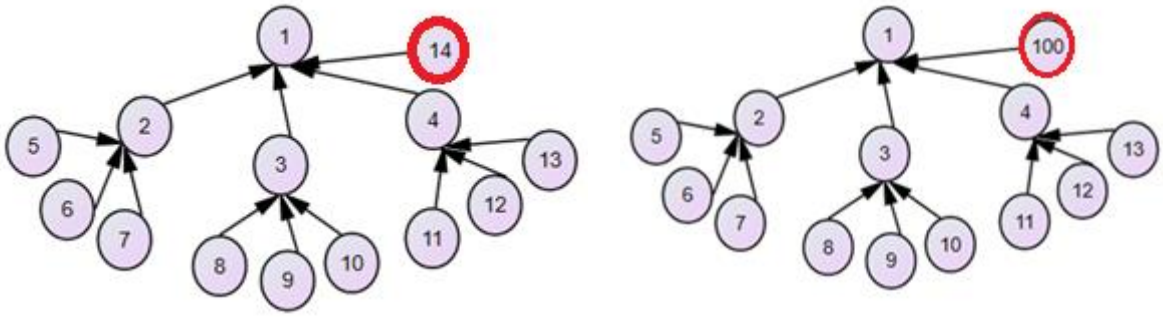
Table 2.2. Number of constructs in each tree.

To diagnose each type of disagreement among the raters in a CAS tree, each perfect tree is compared to a series of 27 possible modifications. Each modification is performed 100 times on each perfect tree. The total number of tests is therefore 16200 (27 x 6 x 100). These modifications are:

- 9 possible direction movements comprising a combination of a movement type (types 1-3) and direction (level, hierarchy, diagonal)
- 3 possible movements where we keep the type constant, and allow random directions

- 3 possible movements where we keep the direction constant, and allow random types
- 3 possible swaps (level, hierarchy, diagonal)
- 8 top and bottom movements, where we restrict one half of a tree. Consider an example with tree *T* which has 5 levels. We first limit movements and swaps for only levels 2 and 3 and then for only levels 4 and 5. It should be noted that by definition, the scores on the top half of the tree will change more than on the bottom half of the tree, given there are fewer nodes on the top half, and thus any change will have a greater effect. However, we wanted to know what the magnitude of the difference would be like.
- 1 random movement/swap where a random change (either one of the 12 movements or 3 swaps) is performed. Each change is equally likely. The aim is to compare the results and evaluate how the statistics change and identify a suitable threshold.

Finally, for the insertion process, we assess our trees by first creating a perfect tree, *T*. Next, we make a copy of the tree, as *T'*. Then for each type of insertion, we randomly add one construct to tree *T* and map it to a node. Contingency table analyses are unable to be performed to compare two different sample sizes. To address this issue, for every missing node in tree *T*, a node is represented in *T'* in the same location with a number not found in *T'*. We repeat this 20 times. As an example, consider a 3-level 3-branch tree as illustrated in Figure 2.8, node 14 is in tree *T* and mapped to node 1. Node 14 does not exist in tree *T'*, hence, for tree *T'*, we insert a dummy node to represent node 14 of tree *T*, thus node 100 is inserted and map to node 1. As this increases the number of branches per node and not the number of levels, we consider this insertion as a level insertion.

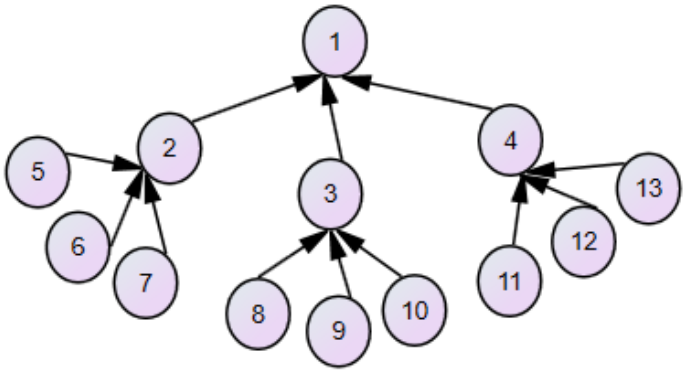


(a) Tree T	(b) Tree T’

Figure 2.8. Level insertion for a 3-level 3-branch tree.

2.3.2 Data collection

Each variation of the tree was represented in a contingency table as follows. First, every node in the tree was given a number from *1* to *N*. 1 was the root node. The tree was then translated into a two-column table. The first column denoted the parent node, and the second column denoted the child. Table 2.3 presents a 3-level 3-branch tree transformed into two columns. As seen, in Table 2.3, there are 13 child constructs and each parent construct has 3 children, each row represents a child and a parent. As an example, row 11 shows that child construct 11 belongs to parent construct 4.



Row	Child	Parent
1	1	0
2	2	1
3	3	1
4	4	1
5	5	2
6	6	2
7	7	2
8	8	3
9	9	3
10	10	3
11	11	4
12	12	4
13	13	4

Table 2.3. Transformed Tree

2.4 Analysis

The perfect tree was placed alongside the modified tree and a statistical comparison between the two trees was performed. Each pair of trees was compared on three statistics, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954). Recall that we analysed 6 possible perfect trees varying in number of levels and branches. Next, for the first 100 runs of each type of movement or swap, the 6 trees are transferred into a table and each child and parent is combined into an individual column and the means and standard deviations of the 3 statistics are calculated. In addition, the mean change (i.e., how much each statistic changes from one run to the next) and standard deviations of the mean change were calculated for the first 100 runs of each movement and swap (a total of 98 mean changes). Lambda, Kappa, and Gamma of the 100 rounds for 6 trees were recorded in each column and a paired sample t-test for each pair of the measures was calculated. Finally, for each type of insertion process, we calculate Lambda, Kappa, and Gamma of the 20 rounds.

2.5 Results

To build suitable thresholds for comparing and assessing CAS trees, we compared each of our hypothetical "perfect" trees to the modified tree and measured the statistic, repeating this process many times. Our results demonstrate Lambda (λ), Kappa (ƙ), and Gamma (γ) change at different rates depending on the kind of movement and swap performed. Table 2.4 presents the summary of these changes. There are several insights for each movement or swap, which we discuss below.

Type of modification	Description	Changes in Lambda (λ), Kappa (ƙ), and Gamma (γ)
Movement types	Modifications that impact descendants	Lambda (λ) decreases at a faster rate in type 2 than other types
Swaps	Modifications that impact the mapping of two constructs	Lambda does not change
Hierarchy movements and swaps	modifications that impact the number of levels in in a tree	Gamma (γ) decreases the fastest
Level movements and swaps	modifications that impact the number of branches in a tree	Kappa (ƙ) decreases the fastest
Diagonal movements	modifications that impact both number of levels and branches in a tree	Lambda (λ) decreases the fastest
Top and bottom movements and swaps	modifications that impact higher levels versus lower levels	Lambda (λ), Kappa (ƙ), and Gamma (γ) decrease faster in top levels than bottom levels
Insertions	Modifications that increase the number of nodes in either branches per node or in the number of levels.	Behaviour of Kappa (ƙ), and Gamma (γ) is similar to level and hierarchy movements

Table 2.4. Summary of changes of Lambda (λ), Kappa (ƙ), and Gamma (γ) for different types of modifications.

2.5.1 Movements

There are several insights for each direction or type of movement. Table 2.5 presents the mean and standard deviations and Cohen's distance for the first 100 runs of each movement for the 6 trees. Cohen's distance provides a measure of the strength of the difference in a t-test (Cohen, 1988). In addition, Table 2.6 presents the mean changes of the measures for each directional movement.

Type of Change	Lambda Mean (SD)	Kappa Mean (SD)	Gamma Mean (SD)	Cohen's Distance for Kappa and Gamma	Paired Samples t-test for Kappa and Gamma df=599)	Sig. (2-tailed)
Hierarchy movement	0.570 (0.125)	0.432 (0.294)	0.2005 (0.472)	0.59	-27.947	<0.001
Hierarchy Movement type 1	0.68 (0.014)	0.723 (0.192)	0.57 (0.189)	0.378	-18.046	<0.001
Hierarchy Movement type 2	0.555 (0.240)	0.411 (0.258)	0.108 (0.461)	0.81	-11.985	<0.001
Hierarchy Movement type 3	0.583 (0.250)	0.388 (0.25)	0.352 (0.242)	-0.193	-4.049	<0.001
Level movement	0.479 (0.283)	0.354 (0.26)	0.522 (0.275)	-0.615	18.97169	<0.001
Level Movement Type 1	0.897 (0.388)	0.874 (0.059)	0.922 (0.0057)	0.826	35.404	<0.001
Level Movement Type 2	0.356 (0.307)	0.304 (0.306)	0.779 (0.2603)	1.6738	29.416	<0.001
Level Movement Type 3	0.671 (0.165)	0.525 (0.25)	0.699 (0.17)	-0.784	16.89	<0.001
Diagonal movement	0.415 (0.251)	0.399 (0.267)	0.358 (0.248)	0.159	-5.557	<0.001
Diagonal Movement Type 1	0.628 (0.112)	0.698 (0.136)	0.744 (0.118)	-0.361	-11.56	<0.001
Diagonal Movement Type 2	0.479 (0.25)	0.369 (0.272)	0.351 (0.234)	0.094	31.238	<0.001
Diagonal Movement type 3	0.50 (0.234)	0.435 (0.293)	0.382 (0.263)	0.183	-10.061	<0.001
Movements	0.512 (0.228)	0.4412 (0.268)	0.4125 (0.254)	0.109	-2.639	0.009
Type 1 movement	0.743 (0.117)	0.428 (0.284)	0.414 (0.264)	0.051	-8.23	<0.001
Type 2 movement	0.448 (0.246)	0.371 (0.267)	0.337 (0.287)	0.123	-5.867	<0.001
Type 3 movement	0.568 (0.215)	0.529 (0.275)	0.553 (0.221)	-0.099	5.172	<0.001

Table 2.5. Means, standard deviation and Cohen's distance for the different movements.

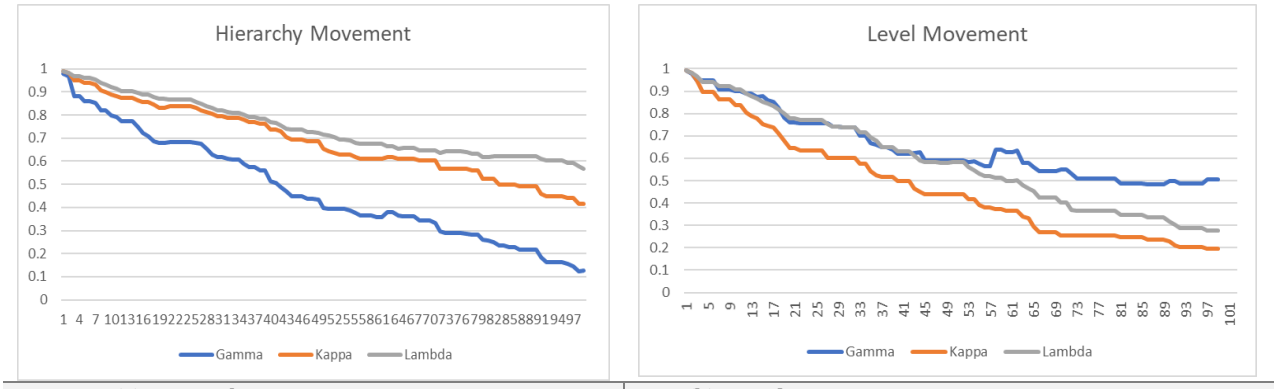
Movement/ Measure	Mean change for Lambda	Mean change for Kappa	Mean change for Gamma
Hierarchy movement	0.000647	0.00088	0.00138
Level movement	0.000660	0.000697	0.000067
Diagonal movement	0.001688	0.001667	0.00163

Table 2.6. Mean changes of different directional movements.

Results indicate that for all hierarchy movements types, Gamma (γ) decreases more dramatically than the other two measures. In addition, the mean for Gamma is lower than the other two measures throughout all types of hierarchy movement. As an example, in a 4-level 3 branch CAS tree as illustrated in Figure 2.9(a), Gamma (γ) in run 20 drops from 0.978 to 0.683 in hierarchy movements while Kappa (ƙ) drops from 0.966 to 0.839. Paired sample t-tests between Gamma and Kappa (the next lowest measure) are all statistically significant.

In level movements, results indicate the mean for Kappa for the 6 trees is lower than the other two measures. In addition, Kappa (ƙ) decreases at the fastest rate of all three measures. All

changes are statistically significant when Kappa is compared to Gamma, the next lowest measure. As an example, in a 4-level 3-branch CAS tree as presented in Figure 2.9 (b), the mean change for Kappa (ƙ) is 0.0036, while Gamma (γ) is only 0.0016 in level movement. In addition, in level movements, for a 4-level 3-branch CAS tree, Kappa (ƙ) in run 20, drops from 0.991 to 0.635, while Gamma (γ) drops from 0.999 to 0.761.



(a) Hierarchy Movement (b) Level Movement

Figure 2.9. Presents the difference between kappa and gamma in level and hierarchy swaps for a 4-level 3-branch tree.

In diagonal movements, Lambda (λ) decreases at a faster rate than any other movement as shown in Table 2.7. As an example, in a 4-level 3branch CAS tree, Lambda (λ) in diagonal movements, in run 20, drops from 0.982 to 0.77, while in level movement it drops from .9871 to 0.8423 and in hierarchy movements it drops from 0.991 to 0.866. We ran a paired sample t-test on 6 different CAS trees to compare the raw scores of Lambda (λ) with the next lowest measure (Kappa (ƙ) or Gamma (γ)) for different CAS trees. The results for each pair was significant which indicates that the measures change at different rates.

Finally, as shown in Table 2.5 in type movements, Lambda (λ) is more sensitive to type 2 movements; the mean for Lambda (λ) is lower compared to other movement types (1 and 3). Type 2 movements consist of two steps, one, move of the parent node and two, move of the child nodes to the former parent's parent node. These two steps have a bigger impact on Lambda than other measures, as more than one node is impacted.

2.5.2 Swaps

Our insights, which are shown in Table 2.7 and 2.8 concerning swaps are as follows:

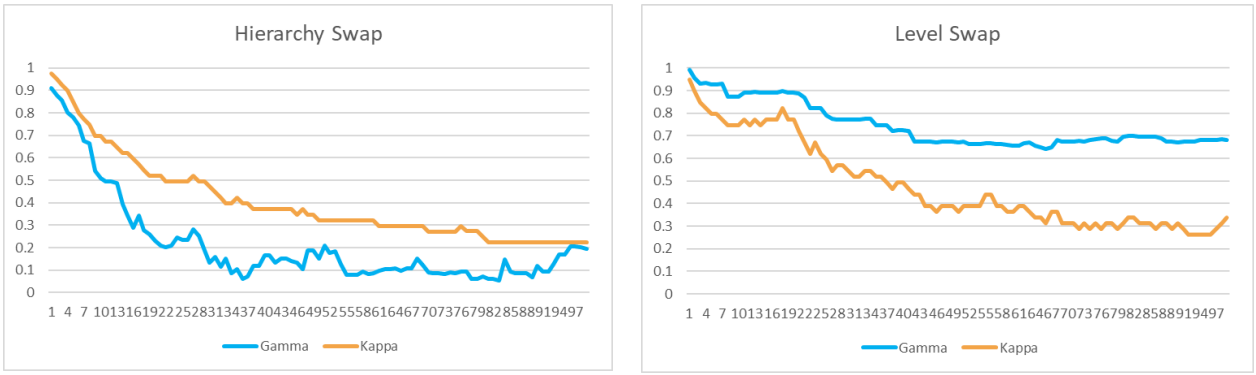
- Lambda does not change in swaps, as both the mean and mean change are zero.
- For hierarchy swaps, the mean of Gamma (γ) is lower than Kappa (ƙ), and mean changes for Gamma (γ) is higher than mean difference for Kappa (ƙ), which indicates that Gamma (γ) drops faster than Kappa (ƙ). In the example shown in Figure 2.10(a), in a 4-level 3-branch tree, Gamma (γ) in run 20, for hierarchy swap drops from 0.9092 to 0.231, while Kappa (ƙ) drops from 0.974 to 0.520. The difference between Kappa and Gamma is statistically significant. This is consistent with Kappa and Gamma's behaviour for hierarchy movements.
- In level swaps, Kappa (ƙ) tends to decrease faster than Gamma (γ) as the mean change of Kappa (ƙ) is higher than Gamma (γ). As an example, as shown in Figure 2.10(b), in a 4-level 3-branch tree, Kappa (ƙ) drops from 0.949 to 0.72 while Gamma (γ) drops from 0.992 to 0.889 in level swaps. The difference between Kappa and Gamma is statistically significant. This is consistent with Kappa and Gamma's behaviour for level movements.

Type of Change	Lambda Mean (SD)	Kappa Mean (SD)	Gamma Mean (SD)	Cohen's Distance for Kappa and Gamma	Paired Samples t-test for Kappa and Gamma (df=599)	Sig. (2-tailed)
Hierarchy swap	1 (0)	0.28 (0.22)	0.213(0.27)	0.191	-5.423	<0.001
Level swap	1 (0)	0.30 (0.23)	0.549 (0.21)	-0.75	35.661	<0.001
Diagonal swap	1 (0)	0.184 (0.24)	0.138 (0.302)	0.124	-7.193	<0.001

Table 2.7. Mean change and standard deviations for the first 100 runs of each swap for the 6 trees.

Type of Change	Mean Difference and (SD) for Lambda	Mean Difference and (SD) for Kappa	Mean Difference and (SD) for Gamma
Hierarchy Swap	0 (0)	0.003 (0.023)	0.007 (0.035)
Level Swap	0 (0)	0.004 (0.029)	0.002 (0.013)
Diagonal Swap	0 (0)	0.014 (0.16)	0.015 (0.08)

Table 2.8. Mean and standard deviations for the first 100 runs of each swap for the 6 trees.



(a) Hierarchy Swap (b) Level Swap

Figure 2.10. Presents the difference between kappa and gamma in level and hierarchy swaps for a 4-level 3-branch tree.

2.5.3 Top and Bottom Movements and Swaps

Given the sample size in a top movement/swap is always smaller than in the equivalent bottom movement/swap, our results unsurprisingly indicated that top movements and swaps decrease at a faster rate than bottom movements and swaps. As an example, Table 2.9 demonstrates the mean changes and standard deviation of top, bottom, and general hierarchy and level movements and swaps for a 4-level 3-branch CAS tree. As presented in Table 2.9, Gamma (γ) decreases the fastest in top hierarchy movements and swaps, as the mean change is higher. In contrast, Kappa (ƙ) decreases the fastest in top level movements and swaps.

Type of hierarchy swap/movement	Mean Difference and (SD) for Lambda	Mean Difference and (SD) for Kappa	Mean Difference and (SD) for Gamma
Top hierarchy swap	---	0.0072(0.024)	0.0074 (0.043)
Bottom hierarchy swap	---	0.005(0.019)	0.0023 (0.02)
Hierarchy Swap	---	0.009(0.012)	0.0072(0.03)
Top hierarchy movement	0.003 (0.005)	0.00501(0.12)	0.00829(0.016)
Bottom hierarchy movement	0.002 (0.004)	0.00419(0.03)	0.0001(0.002)
Hierarchy movement	0.004 (0.004)	0.005(0.09)	0.008(0.013)
Top level swap	----	0.005 (0.019)	0.002 (0.019)
Bottom level swap	----	0.0005(0.0002)	0.0002(0.00025)
level Swap	---	0.006 (0.025)	0.003 (0.012)
Top level movement	0.0042 (0.005)	0.0066 (0.0087)	0.003 (0.004)
Bottom level movement	0.0034 (0.004)	0.003 (0.004)	0.0012 (0.005)
level movement	0.0053 (0.006)	0.006 (0.009)	0.004 (0.012)

Table 2.9. Mean difference and standard deviation for top and bottom hierarchy swaps and movement for a 4-level 3-branch CAS tree in 100 runs.

2.5.4 Insertion Process in CAS trees

Results demonstrate that according to the type of insertion, Lambda, Kappa, and Gamma change differently. As an example, consider Table 2.10 which presents the results of level and hierarchy insertion for a 3-level tree with 3-5 branches. In total, 20 nodes were added to tree *T* and *T'*. In insertion to levels, Kappa is lower than Gamma, while in insertion to hierarchy, Gamma is lower than Kappa. Lambda drops faster in insertion to hierarchy than to level. The difference between Kappa and Gamma is statistically significant. This is consistent with Kappa and Gamma's behaviour for hierarchy and level movements and swaps. In all three cases, Gamma (γ) is lower than Kappa (ƙ) in hierarchy changes, and Kappa (ƙ) is lower than Gamma (γ) in level changes.

number of branches	Type of insertion	Mean and (SD) for Lambda	Mean and (SD) for Kappa	Mean and (SD) for Gamma	Paired Samples t-test for Kappa and Gamma (df=599)	Sig. (2-tailed)
3-branch	Insertion to hierarchy	0.746(0.09)	0.534(0.17)	0.419(0.15)	-15.13	<0.001
3-branch	Insertion to levels	0.975(0.009)	0.56(0.15)	0.63(0.24)	3.2	0.00467
4-branch	Insertion to hierarchy	0.813(0.08)	0.65(14)	0.44(18)	-25.7365	<0.001
4-branch	Insertion to levels	0.978(0.006)	0.66(0.14)	0.79(0.15)	16.37	<0.001
5-branch	Insertion to hierarchy	0.873(0.02)	0.883(0.09)	0.753(0.12)	-12.84	<0.001
5-branch	Insertion to levels	0.98(0.003)	0.74(0.11)	0.86(0.103)	18.30692	<0.001

Table 2.10. Presents the results of level and hierarchy insertion for a 3-level 3-5 branch tree.

2.6 Threshold Results

Many academic disciplines employ threshold values for "satisfactory" levels of inter-rater reliability. For example, the typical threshold for Cronbach's alpha is 0.7 (Nunnally, 1978), and for Cohen's Kappa is 0.7 (Watkins & Pacheco, 2000). We believe suitable thresholds for comparing two CAS trees are when Lambda > 0.7, Kappa > 0.4 and Gamma > 0.3. These thresholds are established for the following reasons:

- It is important to consider all three measures, because each measure signals different kinds of issues. Specifically, changes in Lambda signify movements are occurring, changes in Gamma suggest hierarchical inconsistencies, while changes in Kappa suggest level inconsistencies.
- The three thresholds combined suggest that regardless of sample size, two trees that score above threshold differ in no more than 30% of their nodes. This has been assessed by testing the thresholds in random movements and swaps.

The question remains as to what happens and how efficient are the measures if, (1) only two of the thresholds were used in comparison of all three thresholds and (2) small changes are made to the thresholds. Table 2.11 presents a summary of comparisons where only two of the three thresholds are used. As the table demonstrates, if only two thresholds are applied, it

is possible for two trees to meet threshold when they have substantial differences from each other. For example, if only λ >0.7 and ƙ> 0.4, then it is possible for our 4 level trees to differ by up to 60% of nodes.

Number of levels	Number of nodes	λ >0.7 and ƙ> 0.4	percentage	λ >0.7 and γ > 0.3	percentage	ƙ> 0.4 and γ >0.3	percentage	λ >0.7 and ƙ> 0.4 and γ >0.3	percentage
3 levels	13	3	23.07	3	23.07	6	46.15	3	23.076
4 levels	40	24	60	16	40	16	40	11	27.5
5 levels	121	49	40.49	38	31.40	38	31.4	35	28.92
6 levels	364	120	32.96	115	31.59	111	30.49	100	28.29
Standard deviation			13.5		5.9		6.4		2.29

Table 2.11. Presents a combination of thresholds for 3-branch trees of 3-6 levels

Table 2.12 presents what other thresholds mean when comparing two trees. As an example, consider a 0.1 change of Lambda from 0.7 to either Lambda > 0.6 or Lambda > 0.8 while holding Kappa > 0.4 and Gamma > 0.3. We count the number of modifications for each tree and calculate the percentages. Table 2.12 presents the results of the impact of such 0.1 sized changes with each threshold and the standard deviation of the percentages demonstrates the accuracy of the thresholds for identifying the estimates.

Results indicated that Kappa and Gamma are especially sensitive to changes to their threshold values. As an example, when only Gamma drops from 0.3 to 0.2, the standard deviation for the percentage of modifications is 13.82 while an increase from 0.3 to 0.4, the standard deviation is 1.21. However, when lambda drops from 0.7 to 0.6, the standard deviation of the percentage of modifications is 5.83 and an increase from 0.7 to 0.8, the standard deviations is 5.14. There are several reasons. One, the thresholds are tested in randomized movements and swaps, as Lambda does not change in swaps, hence small changes to Lambda would be less dramatic. Two, in random movements either or both levels and hierarchy of nodes are affected, which results each measure to be more sensitive to small changes, as each measure not only changes with both movements and swaps but changes more dramatically in swaps. Hence, results of

changes in Kappa and Gamma indicate that for finding suitable thresholds, both measures need to be simultaneously adjusted.

Thresholds	**Number of levels**	**3 levels**	**4 levels**	**5 levels**	**6 levels**	**Standard deviation**
	Number of nodes in tree	13	40	121	364	
λ >0.7, ƙ> 0.4, γ >0.3	modifications	3	12	35	104	
	percentage	23.07	30	28.92	28.5	2.6
λ >0.6, ƙ> 0.4, γ>0.3	modifications	6	17	50	111	
	percentage	46.15	42.5	41.32	30.49	5.83
λ >0.8, ƙ> 0.4, γ >0.3	modifications	3	15	36	104	
	percentage	23.07	37.5	29.75	28.57	5.14
λ >0.7, ƙ> 0.3, γ >0.3	modifications	3	23	51	115	
	percentage	23.07	57.5	42.14	31.59	12.84
λ >0.7, ƙ> 0.5, γ >0.3	modifications	3	13	31	102	
	percentage	23.076	32.5	25.61	28.02	3.47
λ >0.7, ƙ> 0.4, γ >0.2	modifications	3	24	49	112	
	percentage	23.07	60	40.49	30.76	13.82
λ >0.7, ƙ> 0.4, γ >0.4	modifications	3	8	27	83	
	percentage	23.07	20	22.31	22.8	1.21

Table 2.12. Presents the impact of thresholds with small changes 3-branch trees of 3-6 levels.

In addition, results indicate that different combinations of the measures can identify different amounts of modification between two trees. Table 2.13 presents four thresholds for when the percentage of modifications are at 5, 20, 25, and 30 percent between two trees. As an example, a threshold of λ >0.75, ƙ> 0.5, and γ >0.4 can identify an estimate of 25% of modifications between two CAS trees, while a threshold of λ >0.85, ƙ> 0.7, γ >0.5 is suitable for identifying an estimate of 15% of modifications of two trees. In addition, Table 2.13 presents less strict thresholds such as a λ >0.65, ƙ> 0.35, and γ >0.25 which can identify an estimate of 40% of modifications between two CAS trees.

Thresholds	Number of levels	3 levels	4 levels	5 levels	6 levels	Standard deviation
	Number of nodes in tree	13	40	121	364	
λ >0.85, ƙ> 0.7, γ >0.5	modifications	2	5	22	53	
15%	percentages	15.38	12.5	18.1	14.5	2.03
λ >0.8, ƙ> 0.65, γ >0.45	modifications	2	8	25	70	
20%	percentages	15.38	20	20.6	19.2	2.04
λ >0.75, ƙ> 0.5, γ >0.4	modifications	3	9	27	87	
25%	percentage	23.07	22.5	22.3	23.9	0.61
λ >0.7, ƙ> 0.4, γ >0.3	modifications	3	12	35	104	
30%	percentage	23.07	30	28.92	28.5	2.6
λ >0.65, ƙ> 0.4, γ >0.25	modifications	4	15	47	121	
35%	percentage	30.76	37.5	38.84	33.241	3.23
λ >0.65, ƙ> 0.35, γ >0.25	modifications	6	17	53	133	
40%	percentage	46.15	42.5	43.8	36.53	3.54

Table 2.13. Presents thresholds according to different amounts of modifications for 3-branch trees of 3-6 levels.

2.7 Diagnosing Problems with Inter-rater Reliability

In addition to providing an estimate of how similar two CAS trees are, the three measures can also be employed to identify the kinds of ways two trees differ. One, if Lambda is low (e.g., <0.7), there is a possibility of cases where there are type 2 movements. This type of movement indicates that while they might agree on the grouping of child nodes, they disagree on the parent of the child nodes.

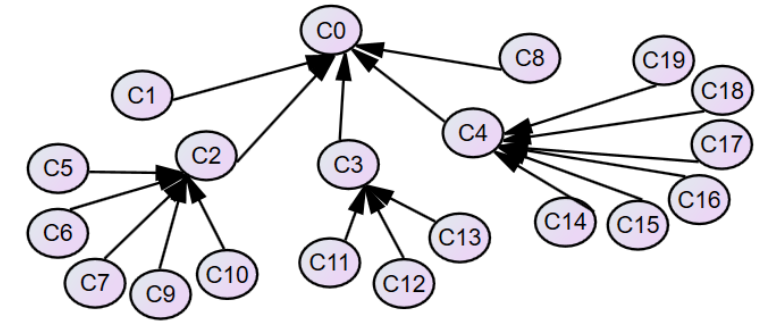
Two, if Gamma (γ) is low (e.g., <0.3), then the principle disagreements are likely with the hierarchy of the constructs. Lambda (λ) is then useful for determining whether the chief problem relates to hierarchy movements (Lambda will be low) or swaps (Lambda will be high).

Three, if Kappa (ƙ) is low (e.g., <0.4), then the principle disagreements are about the mapping of constructs of the same level. Again, Lambda is then useful for determining whether the chief problem relates to level movements or swaps.

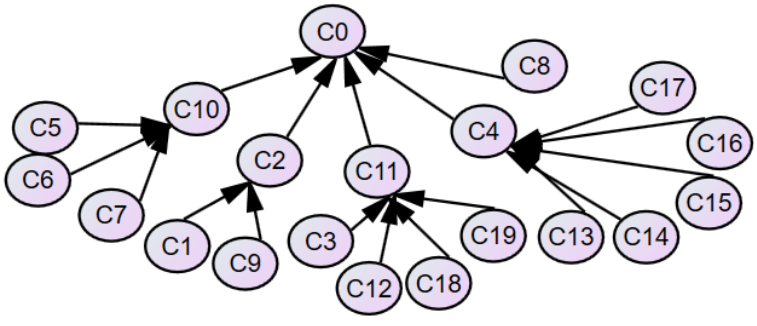
As an example, consider a top-level construct “Price and Value” from a café satisfaction CAS presented in Figure 2.1. Table 2.14 presents the two trees created by the raters (rate 1 and 2) and its transformation to tables and Table 2.15 presents the initial results of the measures. We

have set the thresholds at 30% which suggest there is no more than 30 percent of modification between the two CAS trees. The actual scores for the measures are 0.5 for Lambda, 0.32 for Kappa, and 0.438 for Gamma. The measures provide several insights. One, the measures indicate that the trees are not similar enough, and the problematic constructs will need to be edited accordingly. Two, Kappa being the lowest measure is an indication that the principal problem is the number of disagreements of mapping of constructs of the same level. If we explore the trees, we can see that raters 1 and 2 disagree on the parent constructs of constructs 5, 6, 7,12,13, 18, and 19, which are all located on the same level. Assume we correct this problem so that raters agree on the mapping of constructs 5, 6, 7,12,13, 18, and 19 at this point, the statistics are now, 0.75 for Lambda, 0.745 for Kappa, and 0.807 for Gamma which indicates a strong inter-rater agreement. The combination of the three measures allow us to target the principal problem first. Fixing the principal problem often times can resolve other issues.

CONSTRUCT NUMBER	RATER1	RATER2
CONSTRUCT0	0	0
CONSTRUCT1	0	2
CONSTRUCT2	0	0
CONSTRUCT3	0	11
CONSTRUCT4	0	0
CONSTRUCT5	2	10
CONSTRUCT6	2	10
CONSTRUCT7	2	10
CONSTRUCT8	0	0
CONSTRUCT9	2	2
CONSTRUCT10	2	0
CONSTRUCT11	3	0
CONSTRUCT12	3	11
CONSTRUCT13	3	4
CONSTRUCT14	4	4
CONSTRUCT15	4	4
CONSTRUCT16	4	4
CONSTRUCT17	4	4
CONSTRUCT18	4	11
CONSTRUCT19	4	11



(A)RATER1



(B)RATER2

Table 2.14. Trees and transformed trees of the construct "Price and Value".

Construct	(λ)	(ƙ)	(γ)
Price	0.5	0.32	0.438

Table 2.15. Results of the measures for the construct "Price and Value".

2.8 Limitations

Our analysis reveals several limitations with using lambda, gamma and kappa as measures of CAS trees, first, the thresholds are inapplicable once the number of branches is greater than 7. To demonstrate this limitation, consider different thresholds for 3-level trees with 3-10 branches per node as presented in Table 2.16. Once there are 8 or more branches, the measures are less effective for providing a threshold. As an example, in a 3-level 9 branch per node tree, the thresholds have failed to identify the correct amount of modifications, such as when the thresholds are set to identify 20% of the modifications, they only identify 10%, hence under estimating the amount of modifications. In cases, where one would like to have less strict thresholds or more accurate results, this may be a problem.

In the context of survey research, having too many response options (7 response choices and above) may frustrate or demotivate respondents. When more options are offered, the burden placed on memory is increased (Borgers, Hox, & Sikkel, 2004; Preston & Colman, 2000) and choosing an option may be more challenging. Good CAS tree design should therefore generally avoid having more than 7 branches. Thus, while this is a statistical limitation of our findings, it should not unduly impact people in the field who use CAS tree.

Thresholds	**Number of branches**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**	**Standard deviation**
	Number of nodes in tree	21	31	43	57	73	91	111	
λ >0.85, ƙ> 0.7, γ >0.5	modifications	1	3	6	7	8	8	9	
15%	percentages	4.7	9.6	13.9	12.2	10.95	8.79	8.1	2.77
λ >0.8, ƙ> 0.65, γ >0.45	modifications	2	5	7	11	9	9	11	
20%	percentages	9.5	16.1	16.2	19.3	12.32	9.89	9.9	3.6
λ >0.75, ƙ> 0.5, γ >0.4	modifications	3	6	9	12	11	12	14	
25%	percentage	14.2	19.3	20.9	21.05	15.06	13.18	12.61	3.41
λ >0.7, ƙ> 0.4, γ >0.3	modifications	6	6	10	16	11	12	14	
30%	percentage	28.5	19.3	23.2	28.07	15.06	13.18	12.61	6.26

Table 2.16. Thresholds for different amounts of modifications for 3-level trees with 3-10 number of branches per node.

Two, similar to other studies (van der Ark & van Aert, 2015), we found Gamma (γ) too unstable to provide a reasonable threshold for small samples sizes such as trees with a total number of nodes below 25. However, for trees with a total number of nodes above 25 Gamma (γ) appears to more stable. The point of CAS trees is to facilitate choice between hundreds of options. Thus, for the purposes of CAS trees, Gamma remains a reasonable measure.

Three, due to the exponentially complex computation required, we were unable to run simulations of trees with total number of nodes above 200, we were unable to run the simulations and make any conclusions. However, in our analysis, the measures have been fairly consistent as the growth has been linear as the number of nodes per tree increased. Thus, the threshold results will most likely stay the same in trees with a total number of nodes above 200.

2.9 Conclusion

This study presents an analysis of the use of Lambda, Gamma and Kappa as measures of the similarity of CAS trees and tools for diagnosing their differences. To build suitable thresholds for comparing and assessing CAS trees, we first generated a hypothetical "perfect" tree. We then made a copy of the tree and systematically modified the tree. We created two general types of modifications, movements and swaps.

We repeated the modifications many times and did this for other "perfect" trees of various sizes. We found that:

- Gamma is useful for identifying disagreements with the hierarchy of the constructs.
- Kappa is useful for identifying disagreements on the mapping of constructs of the same level
- Lambda is useful for determining two things, one whether the principal problem is disagreement among single nodes (type 2 movements), which indicates that while raters

agree on grouping of child nodes, they disagree on the parent of the child nodes. Two, a high Lambda in concurrent with a low Kappa or Gamma is useful to detect swaps.

We then proposed thresholds for various levels of inter-rater reliability, as an example, a threshold for when Lambda > 0.7, Kappa > 0.4 and Gamma > 0.3, suggest there is no more than 30 percent of modification between two CAS trees.

This work is particularly useful for assessing the construct and content validity of two CAS trees. Very little research has been done on measuring other types validities for CAS trees. For example, we do not yet have clear techniques for assessing the conclusion validity of CAS, i.e., we don't have a way of knowing if a CAS tree actually works.

Assessing conclusion validity of CAS trees is difficult, because the typical way of assessing conclusion validity is to compare the new instrument with an existing validated instrument (Litwin, 1995) . However, there are few quantitative instruments for measuring why. We have attempted to compare our CAS tree of café satisfaction against blog reviews of cafes. However, we find blog reviews tend not to go into the depth that CAS trees go into. For example, a blog review might state that the food was unappetizing, but not state why it was unappetizing. Our research in this area is ongoing.

References

Anderson, J., & Gerbing, D. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques case I: Sensitivity to data errors. *Journal of the American Statistical Association*, *69*(346), 440–445.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality and Quantity*, *38*(1), 17–33.

Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research. *MIS Quarterly*, *25*(1), 1–16.

Chen, W. (2001). New algorithm for ordered tree-to-tree correction problem. *Journal of Algorithms*, *40*(2), 135–158.

Clauset, A., Moore, C., & Newman, M. (2008). Hierarchical structure and the prediction of

missing links in networks. *Nature*, *453*(May), 98–101.
Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Davis, J. A. (1967). A partial coefficient for Goodman and Kruskal's Gamma. *Journal of the American Statistical Association*, *62*(317), 189–193.
Everitt, B. S. (1992). *The Analysis of Contingency Tables*. *Monographs on Statistics and Applied Probability* (Vol. 45). Chapman & Hall/CRC.
Galesic, M., & 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.
Gefen, D., Straub, D. W., & Boudreau, M.C. (2000). Structural equation modeling and regression : Guidelines for research practice. *Communications of the Association for Information Systems*, *4*(October), 7.
Geoffrion, A. M. (1989). The formal aspects of structured modeling. *Operations Research, 37*(1), 30–51.
Göktaş, A., & İşçi, Ö. (2011). A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. *Metodoloski Zvezki*, *8*(1), 17–37.
Goldreich, O. (2011). Introduction to testing graph properties, *Electronic Colloquium on Computational Complexity Report*, *82*(82), 470–506.
Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764.
Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.
Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & 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).
Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis*. (J. F. Hair, R. E. Anderson, R. L. Tatham, & W. C. Black, Eds.), *International Journal of Pharmaceutics* (Vol. 1). Prentice Hall.
Hambleton, R. K., & Zaal, J. N. (2013). Advances in educational and psychological testing: Theory and applications (Vol.28). *Springer Science & Business Media.*
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.
Heilman, M., & Smith, N. A. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. *The 2010 Annual Conference of the North American Chapter of the ACL*, *3*(June), 1011–1019.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.
Jiang, T., Wang, L., & Zhang, K. (1995). Alignment of trees - an alternative to tree edit. *Theoretical Computer Science*, *143*(1), 137–148.
Klein, P. (1998). Computing the edit-distance between unrooted ordered trees. *In European Symposium on Algorithms*, *Springer B*, 1–102.
Landis, J. R., & Koch, G. G. (1977). The Measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A Comparison of approaches to forming

composite measures in structural equation models. *Organizational Research Methods*, *3*(2), 186–207.
Litwin, M. S. (1995). *How to Measure Survey Reliability and Validity*. Thousand Oaks ; London: Sage.
Nunnally, J. (1987). *Psychometric Theory*. New York: McGraw-Hill.
Porter, S., Whitcomb, M., & Weitzer, W. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, *1*(121), 63–73.
Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15.
Rudick, M. M., Yam, W. H., & Simms, L. J. (2013). Comparing countdown- and IRT-based approaches to computerized adaptive personality testing. *Psychological Assessment*, *25*(3), 769–79.
Sengupta, K., & Te'eni, D. (1993). Cognitive feedback in GDSS: Improving control and convergence. *MIS Quarterly*, *17*(1), 87–113.
Tai, K.-C. (1979). The tree-to-tree correction problem. *Journal of the ACM*, *26*(3), 422–433.
van der Ark, L. A., & van Aert, R. C. M. (2015). Comparing confidence intervals for Goodman and Kruskal's gamma coefficient. *Journal of Statistical Computation and Simulation*, *85*(12), 2491–2505.
Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research : Importance and calculation. *Journal of Behavioral Education*, *10*(4), 205–212.
Wotring, S. C., & Ripley, J. R. (2005). System and method for sharing, mapping, transforming data between relational and hierarchical databases. *Patent No. 6,853,997. Washington, DC: U.S. Patent and Trademark Office*.
Yamakoshi, M., Tanaka, J., & Iwasaki, H. (2016). Security evaluation based on the analytic hierarchy process for first-line anti-counterfeit elements. *Media Watermarking, Security, and Forensics*, 1–6.
You, W., Xia, M., Liu, L., & Liu, D. (2012). Customer knowledge discovery from online reviews. *Electronic Markets*, *22*(3), 131–142.

3 A Method for Developing and Assessing Computer-Adaptive Surveys

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 evaluating the validity of a CAS tree. This paper describes the process of using a variant of q-sorting to validate a CAS tree. In this methodology, trees 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 tree. The trees are then successively transformed to incrementally test if they branch in the same way. The results help researchers not only identify quality items for use in a CAS tree, but also facilitate diagnoses of problems with those items.

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

3.1 Introduction

While traditional survey research has been useful for establishing causal relationships between constructs, it has been poor at "unpacking" constructs to develop a rich understanding of why things are related. As an example, consider a café owner who wants to know not only what aspect of the café (e.g., service quality or food quality) is wrong, but precisely why it is wrong (Fundin & Elg, 2010; Sampson, 1998; Wisner & Corney, 2001). To obtain in-depth information on respondents, a very long survey would be required. Surveys, in general, must strike a balance between trying to understand every respondent's individual views and opinions, and not exhausting the respondent with too many questions (Haschke, Abberger, Wirtz, Bengel, & Baumeister, 2013; Hayes, 1992; Sinclair, Clark, & Dillman, 1993). If the survey is too long, respondents will suffer from a fatigue effect, and not answer questions properly (Berdie, 1989; Deutskens, de Ruyter, Wetzels, & Oosterveld, 2004). If the survey is too short, it will not provide enough information to adequately capture respondent sentiment (Hayes, 1992). 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 & Trochim, 2002).

Computer-Adaptive Surveys (CAS) are most useful for root cause analysis, i.e., explaining why something affects something else. CAS thus offers quantitative positivists a way to approach a domain traditionally held by qualitative researchers, i.e., the ability to answer questions of why. Unlike in a traditional survey, where every item is asked (Hayes, 1992), in a CAS, the previous items determine the next items asked of the respondent. The only items the respondent answers are the ones most salient to the issue being addressed. A CAS can have hundreds of items where each item represents a concept which is mapped into a tree. Items concerning higher level concepts link to items with greater precision. The items then have a parent- child relationship.

CAS validity differs from traditional surveys. One, the items in CAS are arranged in a tree, whereas traditional methods assume a "flat" set of items. Two, respondents legitimately only fill in some questionnaire items- unfilled items cannot be treated as non-responses. Thus, validating a CAS requires different techniques. Unfortunately, methods for validating CAS are under researched- to the point where they are non-existent.

This paper presents a technique we have developed to evaluate the validity of a CAS tree. Specifically, we propose a new q-sorting method, which is applied to a CAS on café satisfaction. In this methodology, trees 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 tree. The trees are then successively transformed to incrementally test if they branch in the same way. The results help researchers not only identify quality items, but also facilitate diagnoses of problems with those items.

The paper is constructed in the following manner. We first introduce CAS in the related literature and describe its design and implementation. Next, we address limitations of traditional methods in construct validity. Following from this, we present a modified q-sorting approach and demonstrate an example of its use on a café satisfaction CAS tree. We then end the paper with a conclusion.

3.2 Computer-Adaptive Surveys (CAS)

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where large constructs are unpacked to explore and identify the different dimensions. Each of these dimensions is represented in an item, where each item in CAS reflects a potential root cause. These items are then mapped in a CAS tree. Items concerning higher level concepts link to items with greater detail. The CAS tree is then transformed into a survey to test which of these items are most relevant to the context of the study. For instance, consider an example in café satisfaction, where all the possible reasons why one would be unhappy with a café is explored and identified and mapped into a CAS tree. The only items the respondent answers are the ones most salient to the issue being addressed. Hence, CAS identifies the root cause of the problem which is perceived by the respondent as most relevant to them. (e.g., which things did you like the least or most). The results of CAS not only indicate which large top-level construct is an issue but precisely which dimension. Their principal advantage is that they allow the survey developer to include a large number of items, where the items are proxies for different possible causes of a relationship. The only items the respondent answers are the ones most salient to the issue being addressed. A CAS is 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 traits by administering questions which are dynamically assigned based on answers to the questions the participant answered previously (Gershon, 2005; Thompson & 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, & 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 & Tymms, 2007), the aim is to understand the reading ability of students to provide better feedback and implement appropriate reading techniques. (Merrell & Tymms, 2007)

While CAT assesses an ability or performance (Hol, Vorst, & Mellenbergh, 2008; Merrell & 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 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 goal of CAS is typically to identify one or a few narrowly defined constructs that explain the relationship between two constructs. For example, a CAS of café satisfaction (i.e., why people are unhappy with a café) would try to determine which of several elements of a café a group of respondents most dislike. The last item 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 main constructs, (1) convenience, (2) service quality, (3) quality of food and drink, (4) price and value, and (5) environment.

CAT relies on potentially complicated Item Response Theory (IRT) functions to determine further questions to ask respondents (Embretson & Reise, 2000; Lord, 1980; Thompson &

Weiss, 2011; Thorpe & Favia, 2012). CAS, in contrast, uses an adaptive version of branching to arrange the items. Low or high scores on a set of items causes the system to retrieve related, but more precise items.

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 the same way. In contrast, in CAS, if the algorithm is designed to identify the construct(s) with the least score, then, as an example, if the first 20 subjects indicated that construct "x" was the least satisfied construct, then to save cost and time, one has the opportunity to direct future respondents to either a different parent construct or have the starting point at a different level. 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 & Weiss, 2009; Ho & Dodd, 2012; Shin, Chien, & Way, 2012) or because further questions in the item bank provide no additionally statistically meaningful information(Thompson & Weiss, 2011). In contrast, a CAS terminates either when one fully traverses a set number of branches of the tree, 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 a relationship, e.g., why are customers unhappy with a café?

CAS can then be suitable for root cause analysis, i.e., explaining why something affects something else. This offers quantitative positivists a way to approach a domain traditionally held by qualitative researchers, i.e., the ability to answer questions of why.

The Root Cause is defined as the primary cause to any problem, which if found and resolved can solve the problem (Dalal & Chhillar, 2013). Root Cause Analysis (RCA) is a method of solving problems from any phase of a process, which affects the project progress or the quality of the overall product. In many cases we would want to know why the problem occurs to not only solve the problem but also for effective problem prevention (Rooney & Van den Heuvel, 2004). There are various well-known methods used for performing RCA such as Cause-Effect Analysis, Brain Storming, Causal Factor Charting, and Fault Tree Analysis (Rooney & Van den Heuvel, 2004).

Like all of these methods, CAS aims to first be comprehensive by identifying all potentially relevant causes. As a method, CAS is, of course, substantively different. Of these methods, CAS is most similar to Cause-Effect Analysis. In Cause-Effect Analysis, a fishbone (Ishikawa) diagram is used to illustrate how various causes can be linked to an identified effect (Bjørnson, Wang, & Arisholm, 2009). CAS similarly arranges potential causes into a structure- in this case, a tree. However, CAS simultaneously applies statistical techniques to identify the most salient cause(s).

3.2.1 How CAS Works

In CAS, respondents perform a depth-first traversal of the tree, where each stage of the traversal involves the respondent rating all items in the stage. Respondents then receive only child items

associated with the lowest or highest rated items. Each respondent could traverse the CAS tree in a different way. To illustrate, see Figure 3.1. If food is the area the customer is least satisfied with, CAS then retrieves items about the quality of the food (i.e., preparation, portion, menu choice). If the customer is least satisfied with menu choice, CAS retrieves items about how the food was cooked, taste, special needs, options, and availability. CAS does not retrieve further items on constructs the respondent rated satisfactorily. As the respondent continues to answer items, CAS navigates deeper down the tree and items roll down until the respondent hits one set of items with no children, for example, that there are insufficient vegetarian items on the menu. If most respondents agree there are insufficient vegetarian items on the menu, this would indicate lack of vegetarian items is the root cause for many customers' dissatisfaction. Of course, not every respondent would navigate the tree the same way. Thus, aggregating the results from respondents allows the researcher to observe the multiple major problems across respondents. In addition, it allows managers of commercial enterprises to quickly find key issues to address.

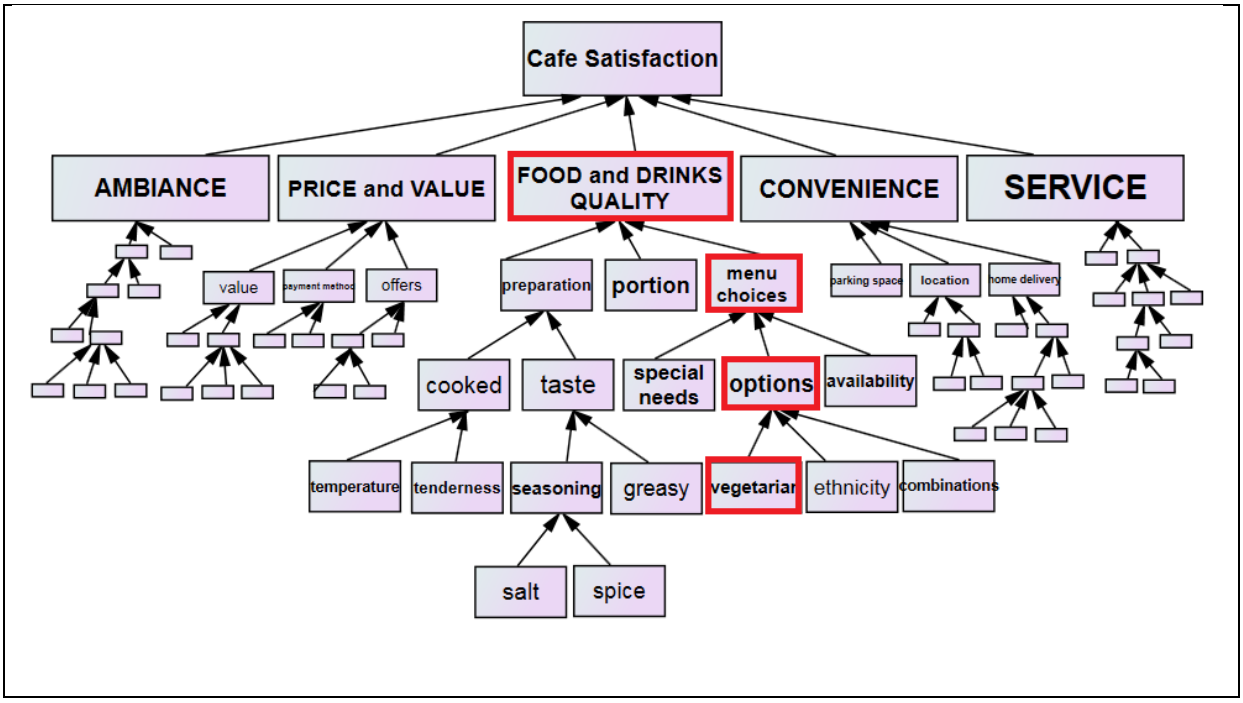


Figure 3.1. How CAS works for café satisfaction.

3.2.2 CAS Tree Properties

CAS trees have particular properties. First, all items in the tree except the penultimate child have at least two children. If an item has only one child, then there is no "branching" and hence no choice; clearly this cannot be allowed. Second, items higher up the tree are formatively defined by items lower down the tree. Hence, items concerning higher level concepts are mapped to items with greater precision. The items thus have a parent- child relationship. This means items higher up in the tree are more important than those lower in the tree. Thus, it is important to first validate the top of the tree, then the next level, etc. This saves effort as there is no point developing subitems for a poorly defined item.

3.3 Assessing the Validity of CAS

Assessing the validity of CAS's results requires different techniques for a number of reasons. First, validation techniques for CAT (e.g.,(Borman et al., 2001)) are not suitable for CAS. As an example, one technique is expert judgement where an expert with content expertise reviews the questions (Hambleton & Zaal, 2013). This method has limited applicability for CAS, because expert judgement is typically about determining whether an item corresponds to a single category of items. While in CAT there are 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) an item properly fits in. In addition, in CAT, construct validity is often determined by comparing the scores obtained from the CAT test against results from a well-accepted, similar non-CAT test (Hambleton & Zaal, 2013; Huff & 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.

Second, popular methods of analysing trees such as tree edit distances which measures the number of changes to convert one tree into another (Tai, 1979) are not suitable for CAS. For one thing, edit distance measures are sample size dependent. An edit distance of 20 is very bad

when comparing two trees with 40 constructs each, but is not so bad if the two trees have over 1000 constructs. In statistical thinking, we want to compare the statistic to some probability distribution to standardize results according to "sample size." We then calculate *p*-values of significance, where the threshold (typically 0.05) is sample size independent. The tree edit distance literature has no equivalent analogue.

Third, in CAS, there are child- parent relationships, where items that are children to other items in the tree should represent some dimension of the parent item. As a result, in CAS, we can expect high correlations between a parent and one child. However, because the subdimensions are orthogonal, we expect low correlations across subdimensions. 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. Hence, traditional statistical construct validity methods such as 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) or structural-equation model-based confirmatory factor analysis (Anderson & Gerbing, 1988; Kline, 1998) have limited applicability.

Finally, constructs in CAS are mapped into a tree, whereas constructs in traditional surveys are assumed to be flat. Q-sorting is a method for assessing construct validity in surveys(Straub & 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 & Grover, 1998). In others, grouping is left to the rater (McKeown & 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, & Rai, 2007).

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

3.4 Types of Errors in a CAS Tree

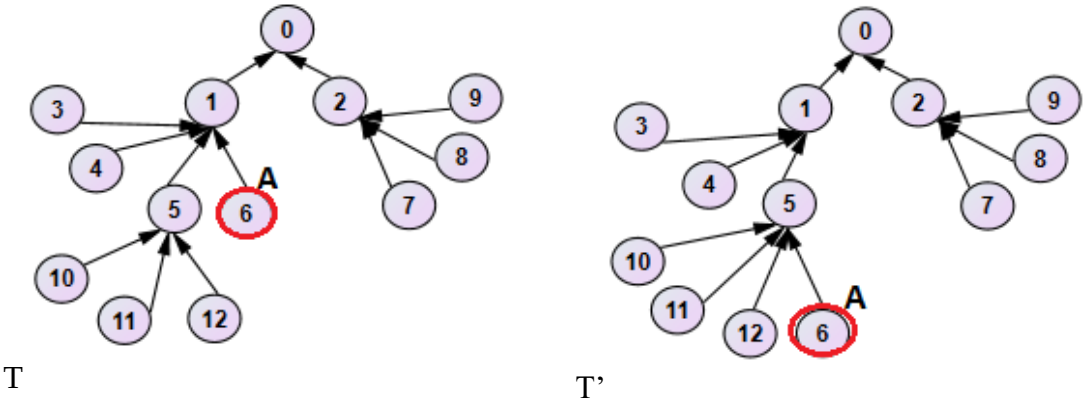
Each CAS tree may contain hundreds of items and those items are mapped based on perception, as each item is mapped based on both its family of items and level in the tree. The family of the item(s) refers to both (1) the parent item and the parent(s) of the parent items, ad infinitum (except for the root), and (2) the child items and the children of the child items, ad infinitum. The level of the tree refers to the distance of the item from the root of the tree. As an example, an item located on the 3rd level is 3 levels below the root and its parent is on the 2nd level. Hence, given identical items, it is possible for two people q-sorting to develop different CAS trees. To develop the final CAS tree, we need to be able to measure the differences between the trees and determine how to fix the differences.

The following are the ways two identical items *A* in CAS trees *T* and *T'* can be arranged differently. In each arrangement, the difference in position of item A represents a certain type of error. Figure 3.2 (a-c) presents these separate arrangements for item A.

First, as Figure 3.2(a) demonstrates, *A* could be found in the same family of items of the tree, but be placed in different levels. In the figure, child items share the same direct and indirect parents. Both are descendants of item 1, however, while in tree *T* item A is directly mapped to item 1, item A in tree *T'* is directly mapped to item 5. We call this a hierarchy movement.

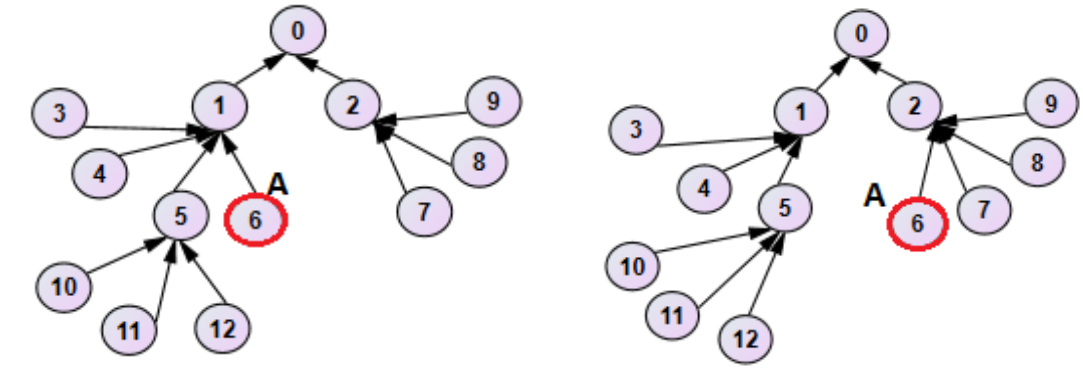
Second (see Figure 3.2(b)), *A* could be from completely different families of items but be found at the same level. In tree *T*, item *A* is a direct child item of item 1, while in tree *T'* it is a direct child of item 2. In both cases, item *A* is on the same level. We call this a level movement.

Third (see Figure 3.2(c)), *A* could be both in a different level and family of item of the tree. We call this a diagonal movement.



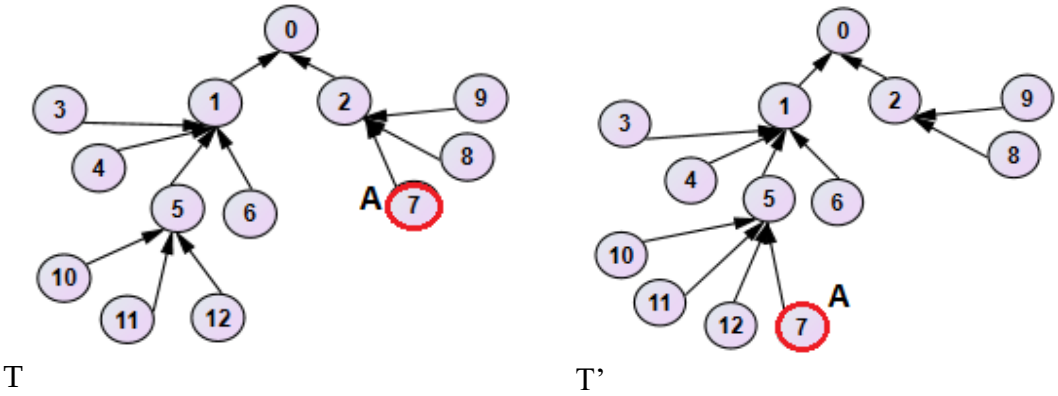
T T’

(a) Hierarchy Movement



T T’

(b) Level Movement



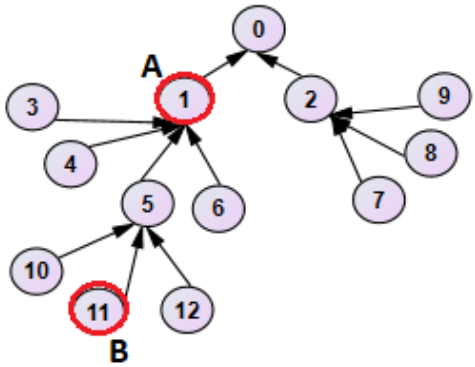
T T’

(c) Diagonal Movement

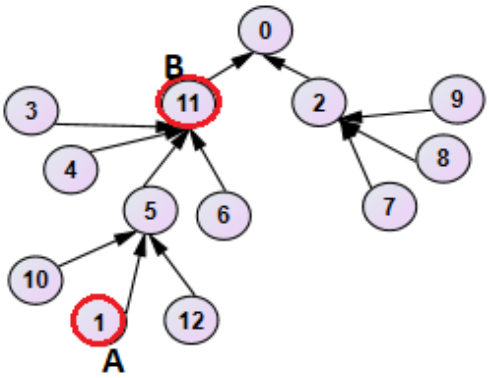
Figure 3.2. Presents problematic item A for CAS tree.

Fourth (see Figure 3.3), there could be another item *B* where *A* and *B* have changed places. We call this a swap. We consider swaps because this reflects a single cognitive difference between two raters rather than two or more cognitive differences. There are effectively three kinds of swaps (hierarchical, level, and diagonal). In the contrasting examples of Figure 3.3, *A* and *B* have swapped places.

Finally, it is possible for one person to be unable to map an item into the tree, while the other was able to do so. In total, there are therefore 7 kinds of errors comprising 3 kinds of movements, 3 kinds of swaps and a situation where one person put an item in the tree, but the other did not.

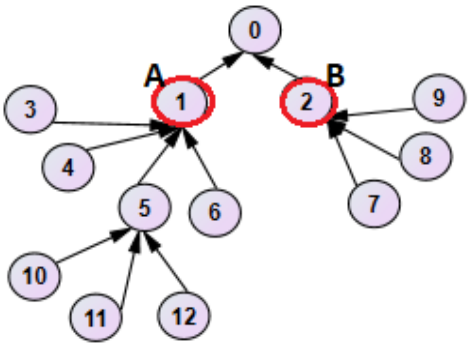


T

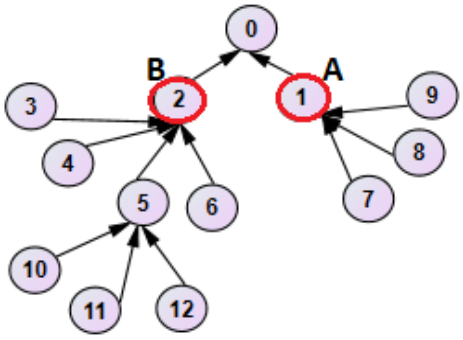


T

(a) Hierarchy Swap

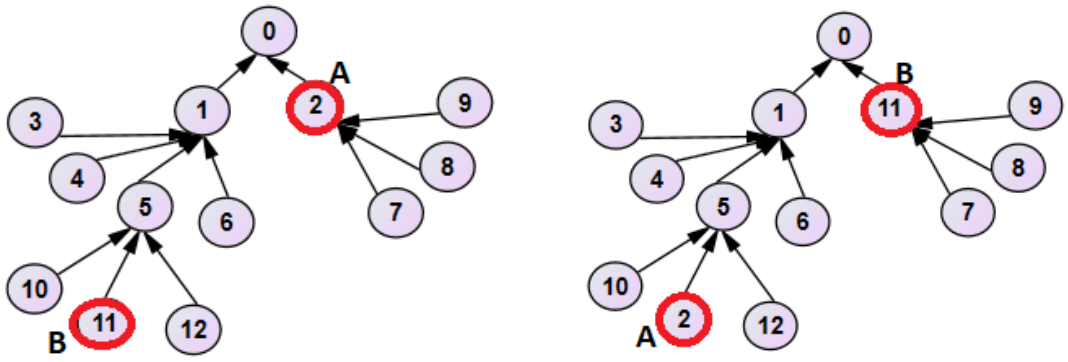


T



T’

(b) Level Swap



(c) Diagonal Swap

Figure 3.3. Presents problematic item A and B for CAS tree.

3.5 Modified Q-Sort Approach

The aim of this modified q-sort is to build and validate a CAS tree. Our technique not only assesses the validity of the tree, but also diagnoses which problem (mapping of items to item) causes problems with inter-rater agreement. Constructs can be viewed as being formed by their items (Bagozzi & Fornell, 1982; Roberts & Thatcher, 2009). The nature and direction of relationships between constructs and items have been discussed in the literature on construct validity and structural equation modeling (Bentler & Chou, 1987; Bollen, 1989). This methodology aims to only validate whether the structure of the CAS tree is aligned and valid. Thus, in this study, traditional theoretical mapping of items to constructs is not discussed. Such mapping are discussed in other research such as (Diamantopoulos, Riefler, & Roth, 2008; Petter et al., 2007). In addition, our technique is similar to other instrument development process (Hoehle & Venkatesh, 2015; Mackenzie, Podsakoff, & Podsakoff, 2011). Our q-sort technique requires three key components:

(1) access to a knowledge sources such as a library, or online forums. The knowledge sources must be rich enough that the items for the CAS tree can be obtained from this knowledge sources. A variety of knowledge sources may be used, such as most documents produced by others like online forums, videos, tutorials, reviews, letters, biographies, speeches, reports, books, (Glaser & Strauss, 1967) and interviews, focus groups, observations. The

knowledge sources will differ depending on the CAS domain and in most cases, a collection of knowledge sources may be required.

(2) at least two raters. Raters must be blind and independent from the study**.** Most studies have indicated that two raters is sufficient (Fleiss, Levin, & Paik, 2013; Gwet, 2008). Raters are that have the following characteristics, one, raters must be an expert in the domain of the CAS tree. Two, raters must have a strong command of the language of the survey. A strong command of the language is necessary because the hierarchical layout of CAS items means raters must understand words that clue a reader into whether an item 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." Three, raters should be able to use certain technology such as digital boards, as use of certain technology increases the efficiency in tree mapping in our q-sort technique.

(3) access to the following pieces of equipment:

- spreadsheet software, such as Microsoft Excel, to be able to transfer the CAS tree into a table, where one column represents the parent items and the other the child items. This is important to analyse the correspondence between the raters' trees.
- statistical software, such as such as SPSS, to calculate Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954). Again, this is important to analyze raters' trees.
- digital boards that will allow one to manipulate and save a tree structure
- non-digital tools such as writing materials (such as paper, cards, or sticky notes), cutting tools (scissors, cutters, or clippers), and writing tools (such as pencil, pen, or marker). The stationery is used to create cards to capture each item in the CAS item bank. Raters will take the cards to physically drop into cardboard boxes in a manner similar to a

traditional q-sort (Block, 1961). The boxes equivalent to the number of top-level constructs in the tree (see Step 2 below).

We have found it is important to have tools capable of both representing CAS trees in a digital (digital boards) and physical (items on paper sorted into boxes) environment Some examples of ideal digital formats are digital whiteboards or large screens. Digital formats allow raters to visualize the overall tree, save their work, and be able to move items around by using tools such as cut and paste. Saving work is important, because raters will often want to restore their work from a prior point. In addition, paper formats are useful as they can be worked on in different locations (i.e., at home, office). It is also useful in locations where internet connectivity is not ideal.

As illustrated in Figure 3.4, our q-sort technique has the following steps; (1) Build items for the CAS, (2) Identify top level constructs for CAS, (3) Incorporate distractors, (4) Select one top-level construct, (5) Sort and map process, (6) Calculate, evaluate and interpret inter-rated scores, (7) Create the final CAS Tree. The components required for each step are identified by color coding the steps. Specifically, blue indicates the use of knowledge sources, green indicates the use of blind and independents raters, and orange means both use of technology and non-digital formats. To illustrate our q-sort technique, we employ an example of developing a CAS to elicit problems customers had with cafes.

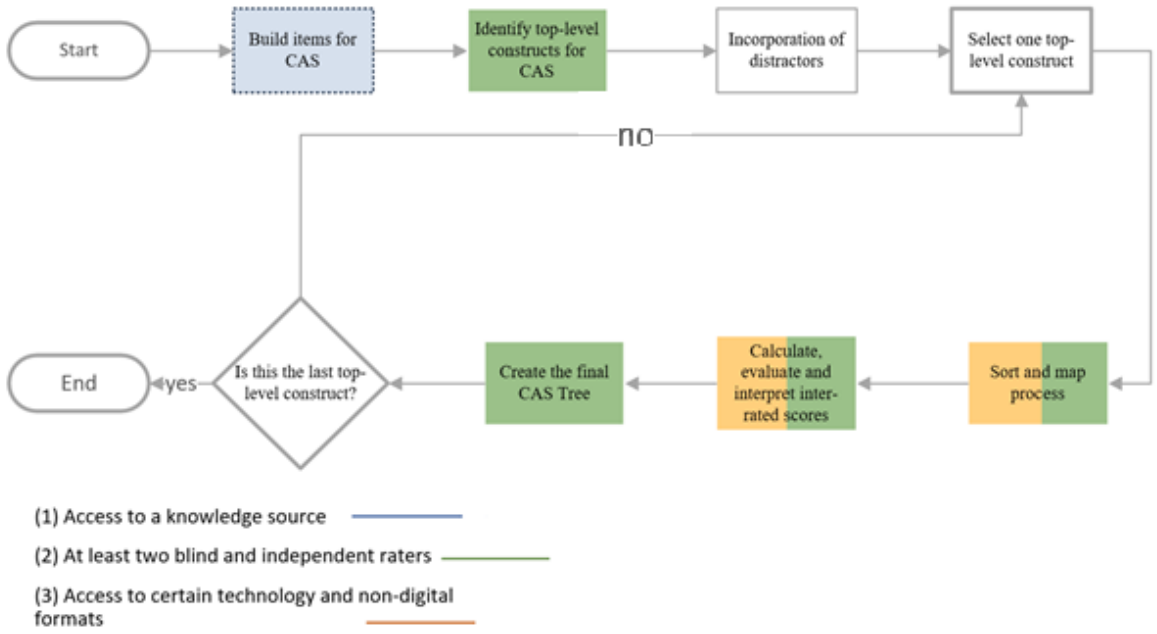


Figure 3.4. Steps for our q-sort technique.

Step 1: Build the items for the CAS. This step is supposed to identify all items that will be employed in the CAS tree. By the end of this step, the items should capture the entire scope of the problem domain of the CAS tree. The concept of scope encompasses both "breadth," i.e., all subdimensions of the concept captured in the CAS tree are represented in the items, and "depth," i.e., for any possible subdimension, all possible actionable causes of that subdimension are captured in the items. Step 1 comprises two substeps which iterate: (a) building the item bank and (b) assessing the item bank. For step (a), we access part of each of our knowledge sources. Each source provides an understanding of possible root causes of the domain. For example, existing journal articles capture information about prior café satisfaction surveys (Hwang & Zhao, 2010; Kim, Moreo, & Yeh, 2005; Liang & Zhang, 2009; Pizam & Ellis, 1999; Pratten, 2004; Ryu & (Shawn) Jang, 2008; Saglik, Gulluce, Kaya, & Ozhan, 2014; Shanka & Taylor, 2005).

Online reviews similarly capture such an understanding, as when there is a sufficient amount of reviews and blogs on social media, a wide range of aspects related to a product or service is discussed in depth (Filieri, 2015; Mudambi & Schuff, 2010). Each of these provides inspiration

to create the items of the item bank. We employ the term item the same way as Edwards and Bagozzi (Edwards & Bagozzi, 2000, p. 156) , i.e., "observable, quantifiable scores obtained through self-report, interview, observation, or other empirical means." Items can be built in various ways, such as, (1) items can be developed from recycling and adapting survey items from previous literature. (2) new items can be built by extracting concepts through observation. A "concept" is a mental image or general notion of something. It summarizes observations and ideas about all the characteristics of that image (Lauffer, 2010, p. 49). For instance, in one online review, one customer stated, "The coffee is substandard (drinkable though), the customer service is non-existent (they yell your coffee orders at your when it's ready) and the food is overpriced and gross." In this context, several concepts exist, including coffee, existence of customer service, employees yelling orders out, price and quality of food. Hence, new items were developed based on those concepts., such as "the quality of coffee of the café was satisfactory" "The quality of food of the café is reasonable" "the customer service of the café was satisfactory", and "the price of food of the café is satisfactory." Next, we assess the face validity (Churchill, 1979) of the items by employing two blind and independent raters. Any item marked by either rater as confusing, is re-edited. As a result of step (a), we collected approximately 400 items for our café' satisfaction survey.

For step (b), we test items for the following issues:

- similarities across items- if two items essentially capture the same idea, one is dropped. We dropped approximately 100 duplicates in our café satisfaction survey, and 300 items remained.
- scope testing- we test for "theoretical saturation" (Bowen, 2008) This is done by ascertaining whether new items were created from step (a). We determine a threshold of a number of times it is required to check for new items (in our café satisfaction instrument, this number was 10), this number is determined by the complexity of the

context of the study and availability and sample size of the knowledge sources. If new items were created in step (a), we reset the number of times back to 0, as there is potential for identifying more items and loop back to step (a). Otherwise, we increment the number of times by 1 (several sources have already been assessed). If the number of times is less than the threshold, we loop back to step (a). Otherwise, we consider that there is no further information from the knowledge sources which we can build our items upon that can be obtained from our knowledge sources and stop

- Content analysis of the items-we test the items for relevance to the context of the study. This is done by employing two independent and blind raters who go through the items, and if both raters mark any item as irrelevant to the context of the study, then that item is dropped. If only one of the raters marks an item as irrelevant, then that item is assessed further in step 2. For example, in our café satisfaction CAS, several items referred to Dutch-style cannabis cafés (coffee shops), often called "Brown Cafes" in the Netherlands (Banys & Cermak, 2016), where marijuana is severed to customers in various forms. However, in our context, use of marijuana in cafes is illegal and thus, not relevant. Hence, these items were mark as not relevant by both raters and dropped. Consequently, in our CAS, nearly 100 items were dropped, leaving 200 items.

Step 2: Identify top-level constructs for CAS. Top-level constructs formatively define the domain of the CAS tree. This step determines what these constructs are, so that all items can be mapped to these constructs. We define the top-level constructs by referring to the literature. For example, café satisfaction traditionally has 5 subconstructs: (1) convenience, (2) service quality, (3) quality of food and drink, (4) price and value, and (5) environment (Kim et al., 2005; Liang & Zhang, 2009). Those five top-level constructs define the breadth of our café satisfaction tree, as there are the total of what defines café satisfaction as there is no other concept that can be mapped to those 5 top-level constructs. To ensure, that no other top-level

construct could be developed, in this step, a traditional q-sort for validating the top-level constructs (Straub & Gefen, 2004) is employed. In the traditional q-sort procedure, each box represents one of these 5 subconstructs. Boxes are labelled with the name of each subconstruct and two independent and blind raters are employed. Our raters independently take every item created in step 1 and place items in the corresponding boxes. We then measure the inter-rater reliability using Cohen's Kappa (Landis & Koch, 1977). If the Cohen's Kappa threshold (typically 0.7) is not met, then there is a problem with the general alignment between the top-level constructs and items. As an example, in our café satisfaction CAS, inter-rater reliability was significant, and kappa was 0.865, above the recommended threshold. In our situation, over 175 items were mapped to top-level constructs., after nearly 20 items were dropped.

If the Cohen's Kappa threshold is met, there are still other issues that need to be resolved. These include:

- Some items may not map to the top-level constructs. When this occurs, a decision has to be made as to whether additional top-level constructs should be established. In this case, a new box is labelled and added to the other boxes. The q-sorting is then repeated and inter-rater reliability is assessed. In our café satisfaction CAS, items which were not mapped to the top-level constructs were dropped. We found all of these irrelevant to our specific context. Thus, from 200 items, nearly 20 items were dropped.

Step 3: Incorporate duplicate and distractor items in the bank. This step is principally employed to partial out error of the CAS instrument from error associated with the rater. The typical CAS test bank can contain hundreds of items. The duplicate and distractor items are used to assess rater attentiveness during the q-sorting technique. It is important to ensure that distractor items are clearly independent of items being assessed by the rater. When raters miss the fact that items are duplicated, or categorize distractor items along with legitimate ones, it signals a lack of rater attention. For example, on a CAS about cafes, one distractor item could

be “The education level is sufficient” as this is clearly independent of the café context. It could be argued that such items are unnecessary, because poor inter-rater reliability would serve as an effective proxy. However, inter-rater reliability is also indicative of poor item phrasing. It is necessary to be able to partial out the effect of raters and items separately, as fatigue is a real concern because of the number of items. As the concern is in regards to the raters identifying the distractors, the number of distractors is subject to the complexity and novelty of the context and number of items a rater would need to sort. In our café satisfaction CAS, the item bank consisted of 175 survey items, to which four duplicate items and distractors were added. Hence a total of 180 items (including the root item) were given to the raters in total.

Step 4: Select one top-level construct at a time. As each CAS tree can contain hundreds of items, to map all of those items at once can be exhausting and too heavy of a cognitive load for raters to perform in a single session. Instead, raters are given only items from one top-level construct at a time. Raters do not sort all items in one session, rather they are given a period of time (e.g., a week) to finish the q-sort technique. This is ok, because the validity of the top-level constructs is already addressed in step 2.

Step 5: Sorting and Mapping items into a tree. In this step, two blind and independent raters who have sufficient knowledge and skills need to be recruited. In our study, we employed raters who had experience in both working and dining in different types of cafes. In addition, prior to study commencement, raters are trained to explicitly draw the tree. 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 tend to perform traditional q-sorts. Thus, in this step two blind and independent raters are told to map items in a tree. Each item is assigned a number from 1 to N, where N is the total number of items. The “root” item, which is a dummy item is developed for statistical purposes is given the number 0

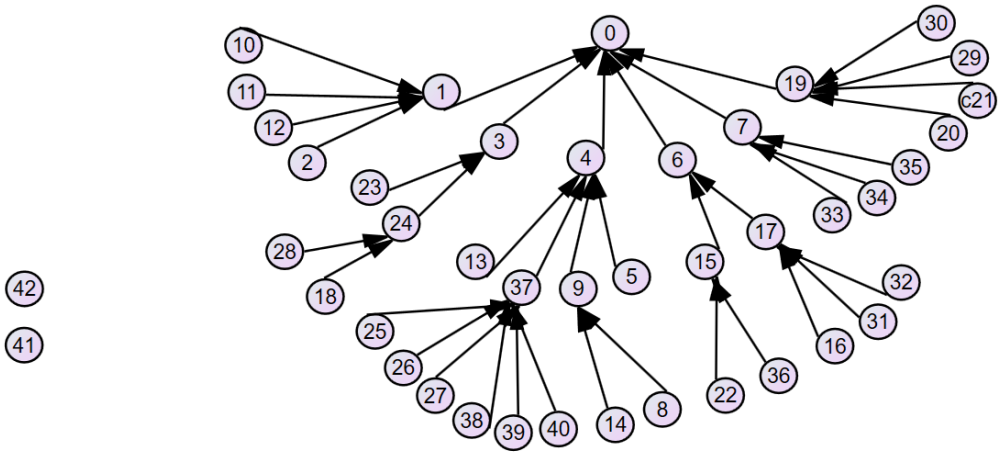
In creating trees, raters must follow certain rules:

- items concerning lower level concepts (children) are mapped to higher level concepts (parents). The root (dummy with value 0) has no parent.
- An item can only have child items if the item has two or more children. This rule should only be enforced on a second attempt at q-sorting.
- A single parent should not have more than 7 child items mapped to it. Prior research argues a good CAS tree design should generally avoid having more than 7 branches, as when too many options are offered, the burden placed on memory is increased (Borgers et al., 2004; Preston & Colman, 2000) which makes choosing the next branch more challenging. However, in cases where a parent item naturally has more than 7 child items, then options such as remapping need to be discussed and agreed upon.
- 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 item.

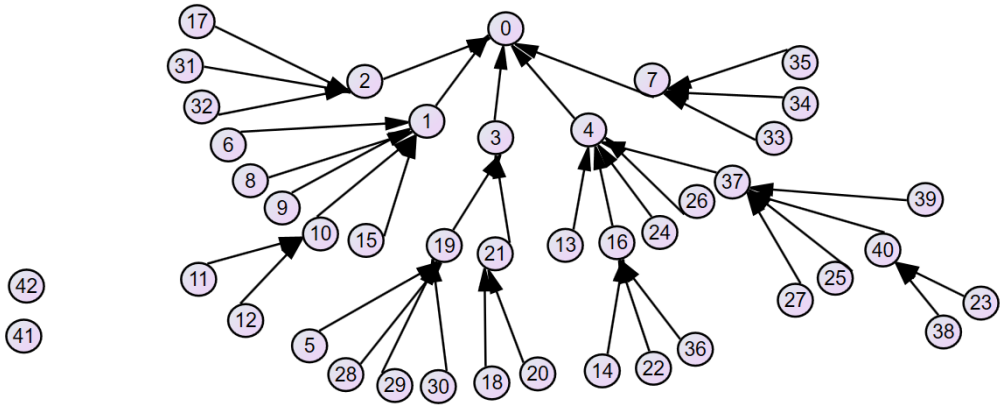
The first round of the q-sort can be insightful in several ways. (1), it is often useful to allow raters to make mistakes to identify potentially problematic items. As an example, if only one child item is mapped to a parent item, this indicates that the parent item is either underdeveloped and future research is required to fully explore and unpack the item, or saturation of that concept has not been complete and further investigation is necessary. (2) It is possible for an item other than the distractor to not be mapped to any item by one or more raters. This indicates one of two possibilities, (a) such item does not belong in the tree because of either bad wording or the item may be vague, or (b) raters have made mistake(s). In either case, further investigation is necessary.

As an example of the process of mapping items into a tree, consider Figure 3.5, which presents the result of two raters developing trees around the concept of "Café Environment." For both raters, Item 1, which refers to the hygiene of the café, is mapped to item 0, which is the top-

level construct of the tree. All top-level constructs are located in level 1. Item 6 refers to the level of comfort of the seating of the café and item 7 refers to staff appearance. For rater 1, item 6 (level of comfort of the seating of the café) and item 7 (staff appearance) are children of item 0 in level 1. Hence, rater 1 views these two items as different families. However, rater 2 disagrees with the mapping of item 6 (level of comfort) and has mapped it to item 1(hygiene of the café). Hence, rater 2 views level of comfort of seating of the café (item 6) as the child of hygiene (item 1). Both raters identified two items (item 41 and 42 for both raters) as unconnected. These items contained the distractor items, which suggests raters were paying attention when performing the q-sort. The two trees created by the raters would then need to be evaluated and the discrepancies resolved, in order to create the final tree. Steps 6 and 7 explain this process further.



(a) Rater 1



(b) Rater 2

Figure 3.5. Raters' tree diagram for the item "Environment".

Step 6: Calculate, evaluate and interpret inter-rated scores. To determine whether the trees are "similar enough" to be considered valid, we perform an iterative contingency table analysis and interpret the results. This step consists of two sub-steps, (1) transformation of the CAS tree into a table and (2) the assessment of the inter-rated scores.

In the first sub-step, in order to calculate the similarity of each tree, certain measures of association are used. To perform these calculations, the trees developed by the raters would need to be transformed into a table. The table has five columns, which are the list of items (1 column), the parent column from each rater, and the level columns for each rater. In the analysis we will compare the parent columns of the two raters using the level columns to restrict the data for analysis. As an example, Table 3.1 demonstrates the transformation of the tree for the construct "Environment" of the café satisfaction CAS. Each rater has mapped the items and created a tree. Items 41and 42 are disconnected from the tree, and hence the mapping is null. Raters may disagree on the level or concept of the items which may create trees with different number of levels. In this case, the tree for rater 1 has 4 levels while the one for rater 2 has 5 levels (including level 0), the results in Table 3.1 correspond to the diagrams in Figure 3.5.

items	Parent items for Rater1	Parent items for Rater2	Level for Rater 1	Level for Rater 2
item 0	0	0	1	1
item1	0	0	2	2
item2	1	0	3	2
item3	0	0	2	2
item4	0	0	2	2
item5	4	19	3	4
item6	0	1	2	3
item7	0	0	2	2
item8	9	1	4	3
item9	4	1	3	3
item10	1	1	3	3
item11	1	10	3	4
item12	1	10	3	4
item13	4	4	3	3
item14	9	16	3	4
item15	6	1	3	3
item16	17	4	4	3
item17	6	2	3	3
item18	24	21	4	4
item19	0	3	3	3
item20	19	21	3	4
item21	19	3	3	3
item22	15	16	4	4
item23	3	40	3	5
item24	3	4	3	3
item25	37	37	4	4
item26	37	4	4	3
item27	37	37	4	4
item28	24	19	4	4
item29	19	19	3	4
item30	19	19	3	4
item31	17	2	4	3
item32	17	2	4	3
item33	7	7	3	3
item34	7	7	3	3
item35	7	7	3	3
item36	15	16	4	4
item37	4	4	3	3
item38	37	40	4	5
item39	37	37	4	4
item40	37	37	4	4
item41	null	null	null	null
item42	null	null	null	null

Table 3.1. Mapping of items to each other for the top-level construct "Environment".

In the second sub-step, as presented in Figure 3.6, the analysis begins at the top of the trees, because disagreement at higher levels of the trees indicate problems with broader, more general items. The parent columns are used to calculate three measures of association. 3 measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954) together provide useful information for evaluating the

similarity of two tress. Prior research recommends thresholds of Lambda>0.7, Kappa >0.4 and Gamma>0.3. This corresponds to about 30% of the two trees being different. Thus, so long as the trees meet satisfactory thresholds for similarity, we add one more level and perform the comparison again. If the measures do not meet satisfactory levels than the scores are interpreted, and the constructs are edited. Editing items can consist of relabelling, deleting, adding of items. The edited items are then given back to the raters to be remapped and the measures are recalculated.

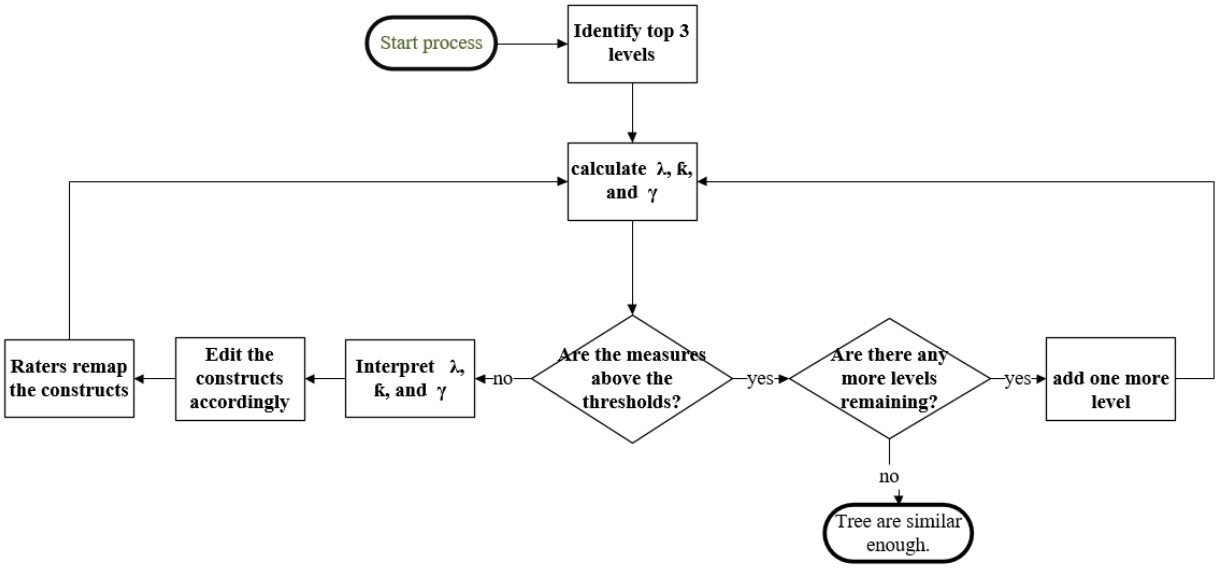


Figure 3.6. Flowchart of the q-sorting evaluation process.

Our analysis is based on prior research which recommends employing three measures, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1954). Each of these three statistics behave differently depending on what is inconsistent between two trees, thus providing both metrics for assessing alignment between two trees developed by independent raters as well as identifying the causes of the differences. As presented in Table 3.2, prior research finds Gamma is particularly sensitive to differences in levels (hierarchy movements or swaps), while Kappa is sensitive to differences within a level of the tree (level movements or swaps). Lambda is

sensitive to “movements” where a single node or branch differs between the two trees, and not sensitive to “swaps” where two nodes or branches are substituted. Prior research argues that employing all three statistics allow the analyst to not only determine the degree of correspondence between two CAS trees, but also how the two trees differ.

Num.	Interpretation of measures	Type of error
1	Gamma is lower than the threshold	Hierarchy movement or swap
2	Kappa is lower than the threshold	Level movement or swap
3	1. Lambda is very low	Movements
4	1. Lambda is higher than the threshold 2. Gamma is low	Hierarchy swap
5	1. Lambda is higher than the threshold 2. Kappa is low	Level swap
6	1. Lambda is higher than the threshold 2. Kappa and Gamma are both low	Diagonal swap

Table 3.2. Summary of interpretation of the measures.

Because raters can be inconsistent, it is possible each rater has put a different number of items at each level. Tests on contingency tables for two different sample sizes can’t be done. To address this issue, if an item doesn’t exist for one rater, we replace the null value that represents the mapping in the table with a number that has not been previously assigned. As an example, consider the two trees created by rater 1 and rater 2 in Figure 3.5. We first identify the top 3 levels which is presented in Table 3.3. Items such as 27 and 28 for both raters are both in level 4 and not assessed at this stage. However, items such as constrict 11 for rater 1 is on level 3 while for rater 2 it is located in level 4. Hence, for these items, we insert dummy items (items 93-100) for rater 2, as presented in Table 3.3. The dummies are italicized in the table. Prior research notes that when this is done, the two trees have the exact same number of items in each level.

Items	Parent items for Rater1	Parent items for Rater2	Level for Rater 1	Level for Rater 2
item 0	0	0	1	1
item1	0	0	2	2
item2	1	0	3	2
item3	0	0	2	2
item4	0	0	2	2
item5	4	*100*	3	3
item6	0	1	2	3
item7	0	0	2	2
item9	4	1	3	3
item10	1	1	3	3
item11	1	*99*	3	3
item12	1	*98*	3	3
item13	4	4	3	3
item14	9	*97*	3	3
item15	6	1	3	3
item17	6	2	3	3
item19	0	3	3	3
item20	19	*96*	3	3
item21	19	3	3	3
item23	3	*95*	3	3
item24	3	4	3	3
item29	19	*94*	3	3
item30	19	*93*	3	3
item33	7	*7*	3	3
item34	7	*7*	3	3
item35	7	*7*	3	3
item37	4	4	3	3

Table 3.3. Insertion of dummy item.

Next, we calculate Lambda (λ) Kappa (ƙ), and Gamma (γ) for the top 3 levels which is tabulated in Table 3.4. In this round, both Lambda (λ) and Kappa (ƙ) are below the threshold while only Gamma is above the threshold as recorded in Table 3.4. Kappa is the lowest score, which identifies that the principal problem is disagreement with the mapping of items of the same level. A scan of the data reveals this to be true- raters disagree on the mapping of items 9, 15, 17, and 21. This indicates that some or all of those items (or their parent items) have problems with either or both the concept or wording. Hence, we edited our items accordingly. As an example, item 9 for rater 1 was mapped to item 4 while for rater 2 it was mapped to item 1. Item 4 refers to the image and appearance of the café and item 1 refers to the hygiene of the café. Item 9 refers to the "scent and smell of the café". The raters, thus disagree on whether the concept of smell refers to image (item 4) or hygiene (item 1). The words scent and smell each

could be interpreted differently, as scent has a relatively positive connection while smell tends to lean on negative. As an example, if one were to say "something smells" it normally means something does not smell nice. Hence, there are two approaches, one, item 9 could be divided into item 9a (smell) and item 9b (scent) of the café, as each item should hold only one concept and repeating the q-sort technique. The other approach would be to drop one of the concepts, as the context of our café satisfaction CAS is more relevant to cafés on university campus. We dropped the concept "scent" as this would be more relevant to upscale cafes or restaurants.

We then recalculated the measures and the scores are above the threshold, as Lambda (λ) is 0.700, Kappa (ƙ) is 0.536, and Gamma (γ) is 0.645, which allow us to proceed to the next level. We added the fourth level and calculated Lambda (λ) Kappa (ƙ), and Gamma (γ). The results indicate that the scores have slightly dropped, and Lambda is now under the threshold. Lambda being below the threshold suggests that there is a possibility of movements. We confirmed this and edited items. As an example, for rater 1, items 11 and 12 are mapped to item 1 while for rater 2 they are mapped to item 10. In addition, item 10 is mapped to item 1 for both raters. Hence, while raters agree on the grouping of the child items (11 and 12) as descendants of item 1, they disagree on whether they are direct children of item 1 or are descendants via question 10. Item 10 refers to cleanliness of the tables, and items 11 and 12 refer to "cleanliness of the utensils" and "cleanliness of the tablecloth." As the concept of cleanliness of the tables is vague and too general, we edited the item to "the cleanliness of the surface of the tables" as cleanliness of the tables could refer to a number of concepts such as uncleared dishes on the table, dirty table cloth, or the dirt on floor underneath the tables. The q-sort technique was then repeated and this time both raters mapped the items 10 (cleanliness of the surface of the tables), 11(cleanliness of the utensils), and 12 (cleanliness of the tablecloth) to item 1 (hygiene of cafe). Finally, we calculated the results for the last remaining level. Table 3.4 presents the final results.

Construct "Environment"	(λ)	(ƙ)	(γ)
three levels	.537	.291	.473
Recalculation of the three levels	.700	.536	.645
4 levels	.647	.447	.637
Recalculation of the 4 levels	.721	.628	.726
All levels for both raters	.706	.628	.721

Table 3.4. Contingency table test of inter-rater agreement for the top-level item "Environment".

Step 7: Building the final CAS tree. When satisfactory levels of significance and suitable thresholds 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, items can be discarded, rewritten, or a final mapping from parent to child item can be agreed upon. As an example, consider Table 3.5 which presents the results for each top-level item for our café satisfaction CAS. In our café satisfaction CAS, we divided the tree in 5 sections and assessed the descendants of each top-level items separately. The measures of the total CAS tree are above each threshold.

Top-level Construct	(λ)	(ƙ)	(γ)
Convenience	.881	.745	.850
Service quality	.700	.546	.506
Price and Value	.704	.487	.562
Environment	.706	.628	.721
Food Quality	.783	.608	.690
Overall for CAS	.760	.641	.669

Table 3.5. Contingency table test of inter-rater agreement for café satisfaction CAS.

However, the raters still disagreed on the mapping of certain items as the scores are not a perfect 1. Hence, the discrepancies need to be discussed, prior to the construction of the final tree. As an example, consider the construct "Environment" from the café satisfaction CAS as presented in Figure 3.7. After the reconciliation, several items were remapped. As an example, raters disagreed on the mapping of items 16 and 17, which refer to the concepts lighting, and crowdedness of the café. The original item for item 6 was "level of comfort of the seating of the café is satisfactory," however, the concept of comfort of seating did not fully encompass the concept of comfort. Hence, we introduced a new item "level of comfort of the café". We

then mapped items 16, 17 and item 6 to the new item 43. Hence, here the item "level of comfort of the café" unpacks to concepts such as lighting, seating, crowdedness, and temperature.

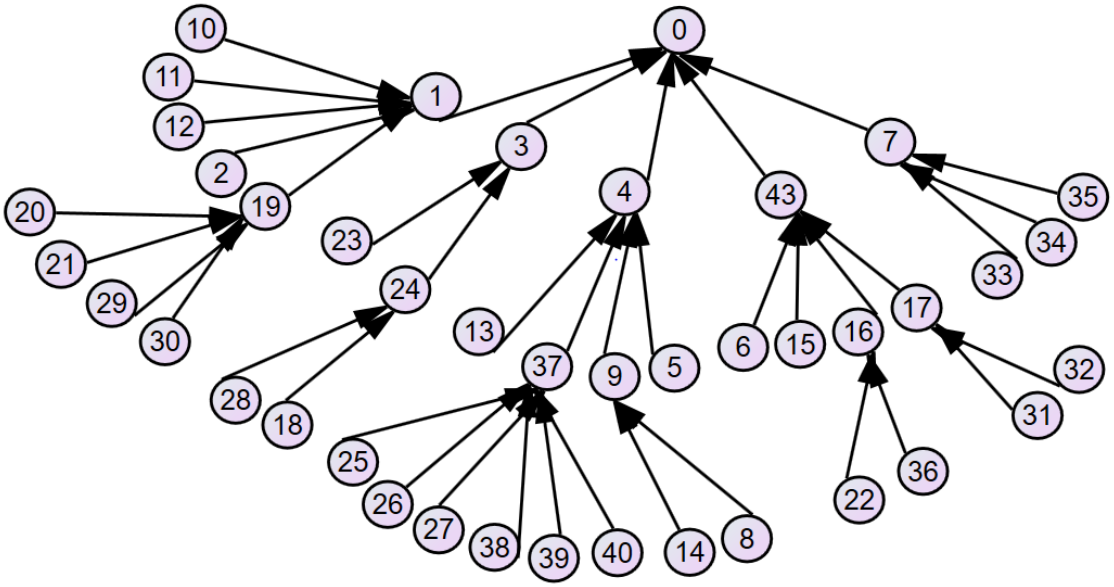


Figure 3.7. Final mapping of the top-level construct "Environment".

3.6 CONCLUSION

This study presents a variant q-sort technique designed to evaluate the validity of CAS trees. CAS items have certain characteristics, such as being multi-dimensional and containing items with parent-child relationships, hence, traditional methods are not suitable. It is important to validate CAS trees, as the child-parent relationships need to be assessed. In our variant q-sort technique, trees 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 tree. The trees are then successively transformed to test if they branch in the same way. Dummy items 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 tree we designed, but also what those problems are.

CAS trees are useful for two things. One, CAS can be used for root cause analysis. CAS drills down to the most relevant cause among respondents to unpack items to develop a rich understanding of why things are related. The results of CAS can be applied to both academic and practice research settings. For academic research, CAS can be applied to answer questions of why (Gregor, 2006). Existing surveys principally are good at establishing correlations between items. Why correlations exist between items is often unknown. CAS allows one to design surveys to answer such questions. Our CAS survey reflects a practical survey, as it allows managers to prioritise issues to address first and optimize their productivity.

Two, our q-sort technique can be useful for approaching research from new directions. Trees are among the most common and well-studied combinatorial structures in many areas such as computer science (Bille, 2005). In addition, taxonomists have become increasingly interested in the theory and practice of comparing tree-like structures (Day, 1985). However, there are few tools and techniques to scientifically analyse trees. This study provides a small step in building such science.

As future research, we hope to explore and assess CAS in several areas. One is to assess CAS in other fields. This study compared a café satisfaction Computer-Adaptive Surveys (CAS) with traditional surveys of the same item bank. We chose café satisfaction instead of a more salient IS topic (e.g., TAM (Bagozzi, 2007; Gefen, Karahanna, & Straub, 2003; Liu, Chen, Sun, Wible, & Kuo, 2010)) because of the availability of existing, relevant instruments. If we were to develop our own TAM/UTUAT instrument, the potential quality of the questionnaire items would be a confound. By using existing items employed in traditional surveys, this confound is eliminated. As future work, we intend to apply CAS to TAM/UTUAT to determine why the constructs encourage intention to use.

Two, we intend to continue developing techniques, strategies, and algorithms to increase the accuracy and efficiency of CAS. CAS uses an adaptive version of branching for respondents to move from one set of items to another set according to a pre-defined criterion. Future study aims to assess and explore other possible options for determining the next set of question(s). Work in this area is ongoing.

References

Anderson, J., & Gerbing, D. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423.

Babcock, B., & 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*.

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, *8*(4), 244–254.

Bagozzi, R. P., & Fornell, C. (1982). Theoretical concepts, measurements, and meaning. *A Second Generation of Multivariate Analysis*, *2*(2), 5–23.

Banys, P., & Cermak, T. L. (2016). Marijuana legalization in California: Rational implementation of the Adult Use of Marijuana Act (AUMA). *Journal of Psychoactive Drugs*, *48*(1), 63–65.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117.

Berdie, D. R. (1989). Reassessing the value of high response rates to mail surveys. *Marketing Research*, *1*(September), 52–65.

Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical Computer Science*, *337*(1–3), 217–239.

Bjørnson, F. O., Wang, A. I., & Arisholm, E. (2009). Improving the effectiveness of root cause analysis in post mortem analysis: A controlled experiment. *Information and Software Technology*, *51*(1), 150–161.

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.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303–316.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality and Quantity*, *38*(1), 17–33.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & 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*, 86, 965–973.

Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research*, *8*(1), 137–152.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64.

Dalal, S., & Chhillar, R. S. (2013). Empirical study of root cause analysis of software failure. *ACM SIGSOFT Software Engineering Notes*, *38*(4), 1.

Day, W. H. E. (1985). Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, *2*(1), 7–28.

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*(1), 21–36.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. *Quality of Life Research*, *4*(3), 1–371.

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research*, *68*(6), 1261–1270.

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

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

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated mode. *MIS Quarterly*, *27*(1), 51–90.

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

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. *Observations*, 1(4).

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611–642.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis*. Prentice Hall.

Hambleton, R. K., & Zaal, J. N. (2013). *Advances in educational and psychological testing: Theory and applications*. Springer Science & Business Media.

Haschke, A., Abberger, B., Wirtz, M., Bengel, J., & Baumeister, H. (2013). Development of short form questionnaires for the assessment of work capacity in cardiovascular rehabilitation patients. *International Journal of Occupational Medicine and Environmental Health*, *26*(5), 742–50.

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

Ho, T.-H., & Dodd, B. G. (2012). Item selection and ability estimation procedures for a mixed-format adaptive test. *Applied Measurement in Education*, *25*(4), 305–326.

Hoehle, H., & Venkatesh, V. (2015). Mobile Application Usability: Conceptualization and Instrument development. *MIS Quarterly*, *39*(2), 1–12.

Hol, a. M., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized adaptive testing of personality traits. *Journal of Psychology*, *216*(1), 12–21.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement, Issues and Practice*, *20*(3), 16–25.
Hwang, J., & Zhao, J. (2010). Factors influencing customer satisfaction or dissatisfaction in restaurtant business using answertree methodology. *Journal of Quality Assurance in Hostpitality & Tourism*, *11*(2), 93–110.
Jackson, K. M., & 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.
Kim, Ay.-S., Moreo, P. J., & Yeh, R. J. M. (2005). Customers' satisfaction factors regarding university food court service. *Journal of Foodservice Business Research*, *7*(4), 97–110.
Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling* (Third). New York: The Guilford Press.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
Lauffer, A. (2010). *Understanding Your Social Agency*. (3rd ed.). London: SAGE Publications Ltd.
Liang, X., & Zhang, S. (2009). Investigation of customer satisfaction in student food service. *International Journal of Quality and Service Sciences*, *1*(1), 113–124.
Liu, I. F., Chen, M. C., Sun, Y. S., Wible, D., & Kuo, C. H. (2010). Extending the TAM model to explore the factors that affect intention to use an online learning community. *Computers and Education*, *54*(2), 600–610.
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., & 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.
McKeown, B. F., & Thomas, D. B. (1988). *Q Methodology (Quantitative Applications in the Social Sciences series, Vol. 66)*. Newbury Park, CA: Sage.
Merrell, C., & Tymms, P. (2007). Identifying reading problems with computer-adaptive assessments. *Journal of Computer Assisted Learning*, *23*, 27–35.
Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly*, *34*(1), 185–200.
Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, *31*(4), 623–656.
Pizam, A., & Ellis, T. (1999). Customer satisfaction and its measurement in hospitality enterprises. *International Journal of Contemporary Hospitality Management*, *11*(7), 326–339.
Pratten, J. D. (2004). Customer satisfaction and waiting staff. *International Journal of Contemporary Hospitality Management*, *16*(6), 385–388.
Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15.
Roberts, N., & Thatcher, J. (2009). Conceptualizing and testing formative constructs. *ACM SIGMIS Database*, *40*(3), 9.
Ryu, K., & (Shawn) Jang, S. (2008). DINESCAPE: A scale for customers' perception of dining environments. *Journal of Foodservice Business Research*, *11*(1), 2–22.
Saglik, E., Gulluce, A. C., Kaya, U., & Ozhan, K. C. (2014). Service quality and customer satisfaction relationship : A research in Erzurum Ataturk University. *American International Journal of Contemporary Research*, *4*(1), 100–117.

Sampson, S. E. (1998). Gathering customer feedback via the internet: Instruments and prospects. *Industrial Management & Data Systems*, *98*(2), 71–82.
Segars, A. H., & Grover, V. (1998). Strategic information systems planning success: An investigation of the construct and its measurement. *MIS Quarterly*, *22*(2), 139–163.
Shanka, T., & Taylor, R. (2005). Assessment of university campus café service: The students' perceptions. *Asia Pacific Journal of Tourism Research*, *10*(March), 329–340.
Shin, C. D., Chien, Y., & Way, W. D. (2012). *A Comparison of Three Content Balancing Methods for Fixed and Variable Length Computerized Adaptive Tests*.
Sinclair, M. D., Clark, J. R., & Dillman, D. A. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rate for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*, 289–304.
Straub, D., & 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., & Bridgeman, B. (2006). Test takers' attitudes and beliefs about the Graduate Management Admission Test. *International Journal of Testing*, *6*(3), 255–268.
Tai, K.-C. (1979). The tree-to-tree correction problem. *Journal of the Association for Computing Machinery*, *26*(3), 422–433.
Thompson, N., & Weiss, D. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Education*, *16*(1), 1–9.
Thorpe, G. L., & Favia, A. (2012). Data analysis using item response theory methodology : An introduction to selected programs and applications. *Psychology Faculty Scholarship*, *20*.
Wisner, J. D., & Corney, W. J. (2001). Comparing practices for capturing bank customer feedback. *Benchmarking: An International Journal*, *8*(3), 240–250.

4 Quantitative Instrumentation for Answering "Why" Questions: AN Empirical Test

Abstract

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. Their principal advantage is they allow the survey developer to input a large number of potential causes. Respondents then roll down through the causes to identify the one or few significant causes impacting a correlation. This paper attempts to demonstrate that CAS are superior instruments for diagnosing root cause. To do this, we compare the conclusion validity of a CAS about café customer satisfaction against two other techniques, specifically (1) a psychometric survey, and (2) a content analysis of online customer reviews. We find that CAS has several advantages over the psychometric survey, as CAS has a higher response rate, requires fewer items for respondents to answer, which reduces fatigue effect, has better item discrimination in that respondents tend to provide more extreme scores in CAS, and has a higher agreement among respondents for each item. In addition, our results suggest CAS is a more robust approach to assessing café satisfaction than online reviews.

Keywords: *Computer-Adaptive Surveys, psychometric surveys, online reviews, conclusion validity*

4.1 Introduction

One fundamental criticism of quantitative methodologies is they are poor at determining why things are the way they are. For example, while psychometric surveys are excellent instruments for establishing the correlational relationship between two variables/constructs, they are unable to unpack their variables/constructs to answer questions of why (Pinsonneault and Kraemer 1993). Computer-Adaptive Surveys (CAS) are claimed to be useful for understanding why two

constructs are correlated. However, the ability and efficacy of CAS (i.e., conclusion validity) for actually unpacking variables/constructs to answer questions of why remains unknown.

This paper attempts to validate the conclusion validity of CAS. To do this, we compare the conclusion validity of a CAS about café customer satisfaction against two other techniques, specifically (1) a psychometric survey, and (2) a content analysis of online customer reviews. We find that CAS has several advantages over the psychometric survey, as CAS has a higher response rate, requires fewer items for respondents to answer, which reduces fatigue effect, has better item discrimination in that respondents tend to provide more extreme scores in CAS, and has a higher agreement among respondents for each item.

We also find a high correspondence between top level constructs of our CAS and online reviews. However, CAS allows one to make more in-depth conclusions about customer dissatisfaction as compared to online reviews. This is because online reviews do not go into as much detail about customer satisfaction as CAS.

The paper is constructed as follows. The next section introduces the related literature, describing CAS and its design. We then present the results of our first study which compares CAS with a psychometric survey of the same item bank. We follow by presenting results of our second study which compares CAS against online review websites. Finally, we discuss the findings.

4.2 Computer-Adaptive Surveys (CAS)

Psychometric surveys and Computer-Adaptive Surveys (CAS) differ, as psychometric surveys are about establishing a correlation between two or more constructs, whereas CAS investigates why those constructs are correlated (i.e., identify root cause). As an example, a psychometric survey establishes a high negative correlation between food quality and satisfaction. A CAS is

used to identify whether customers are unhappy with the taste, temperature, texture, arrangement, etc. of the food.

To identify root cause, it is often necessary to ask many questions. Most surveys are not designed to be very long. Therefore, they are not especially designed to be informative or diagnostic for identifying root cause (Goodman, Broetzmann, & Adamson, 1992; Hayes, 1992; Peterson & Wilson, 1992). For example, most customer satisfaction surveys comprise 30 items or less- because of a lack of respondent patience, often only a single question is asked per construct (Heerwegh & Loosveldt, 2006). A typical CAS will contain at least a hundred questions. Also, in most surveys, the respondent is intended to answer the majority of questions. With a large survey, the respondent is likely to encounter fatigue and quit before providing critical information (Galesic & Bosnjak, 2009; Groves, 2006; Groves et al., 2004; Heerwegh & Loosveldt, 2006; Porter et al., 2004). Finally, the effort to complete a CAS grows logarithmically with the length of the CAS. In contrast, effort grows linearly with the length of a psychometric survey. This is because questions in a CAS are represented in a tree and the respondent navigates down a branch of the tree instead of doing the whole survey.

Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where questions asked of respondents depend on the previous questions asked. Its principal advantage is it allows the survey developer to include a large number of questions. The only questions the respondent answers are the ones most salient to the issue being addressed- in our case, the things about the café the respondent is least satisfied with. In contrast, if the same number of questions were asked on a psychometric survey, the respondent is likely to encounter fatigue and quit before providing critical information (Galesic & Bosnjak, 2009; Groves, 2006; Groves et al., 2004; Heerwegh & Loosveldt, 2006; Porter et al., 2004).

Compared to traditional psychometric surveys, CAS is more focused on absolute differences between factors, instead of relative differences. In other words, the question being asked in the

café satisfaction context is not how much more does ambience affect satisfaction compared to food quality, but what specific aspect of food quality or ambience has the strongest impact on café satisfaction. This is a crucial distinction that has implications for how research is done and what sort of conclusions can be drawn up

One, Computer-Adaptive Surveys (CAS) are multi-dimensional instruments where large constructs are unpacked to explore and identify the different dimensions. Each of these dimensions is represented in an item, where each item in CAS reflects a potential root cause. These items are then mapped in a CAS tree. CAS use a very large item bank and rely on the idea that the response of one item directs the next item. CAS aims to identify the child construct (s) that are perceived by the respondent as most relevant to them.

Two, CAS uses an adaptive version of branching to arrange the questions. The lowest or highest score on a set of questions triggers the retrieval of related, but more precise questions. In addition, the question structures are also different.

Finally, initiation and termination in CAS function in specific ways. In CAS, we could have the first 20 respondents taking a CAS begin with generic questions about the café. If we realise that most respondents are indicating issues with the service, the next 20 respondents might begin at a lower level of the tree- on the service-related questions. A CAS ends either when one has fully traversed a set number of branches of the tree, or when the user reaches some threshold for a proxy for fatigue (e.g., user answers a certain number of questions).

4.3 CAS Implementation and Design

In CAS, respondents perform a depth-first traversal of the tree, where each stage of the traversal involves the respondent rating all items in the stage. Respondents then receive only child constructs associated with the lowest or highest rated constructs. Each respondent could traverse the CAS tree in a different way. To illustrate, see the customer satisfaction CAS

example in Figure 4.1. If food is the area the customer is least satisfied with, CAS then retrieves questions about the quality of the food (i.e., preparation, portion, menu choice). If the customer is least satisfied with menu choice, CAS retrieves questions about how the food was cooked, taste, special needs, options, and availability. CAS does not retrieve further questions on constructs the respondent rated satisfactorily. As the respondent continues to answer questions, CAS navigates deeper down the tree and questions roll down until the respondent hits one set of constructs with no children, for example, that there are insufficient vegetarian items on the menu. If most respondents agree there are insufficient vegetarian items on the menu, this would indicate lack of vegetarian items is the root cause for many customers' lack of satisfaction . Of course, not every respondent would navigate the tree the same way. Thus, aggregating the results from respondents allows the researcher/ manager to observe the multiple major problems across respondents.

However, it is possible there are more candidate constructs in that level of the hierarchy than allowed by a threshold. For example, assume initially that the respondent rated question1 "I am overall satisfied with the ambiance of the café" with a 1, and question3 "I am overall satisfied with the quality of food/drinks of the café" with a 1. CAS must retrieve constructs associated with the least scoring question. But CAS cannot determine which of the questions 1 or 3, to choose from. At this point, CAS asks the respondent which of questions 1 or 3 is most relevant to the respondent. If question 3 is chosen, then the constructs mapping to question 3 are retrieved by CAS. The process repeats until no subsidiary constructs can be selected, whereupon CAS stops.

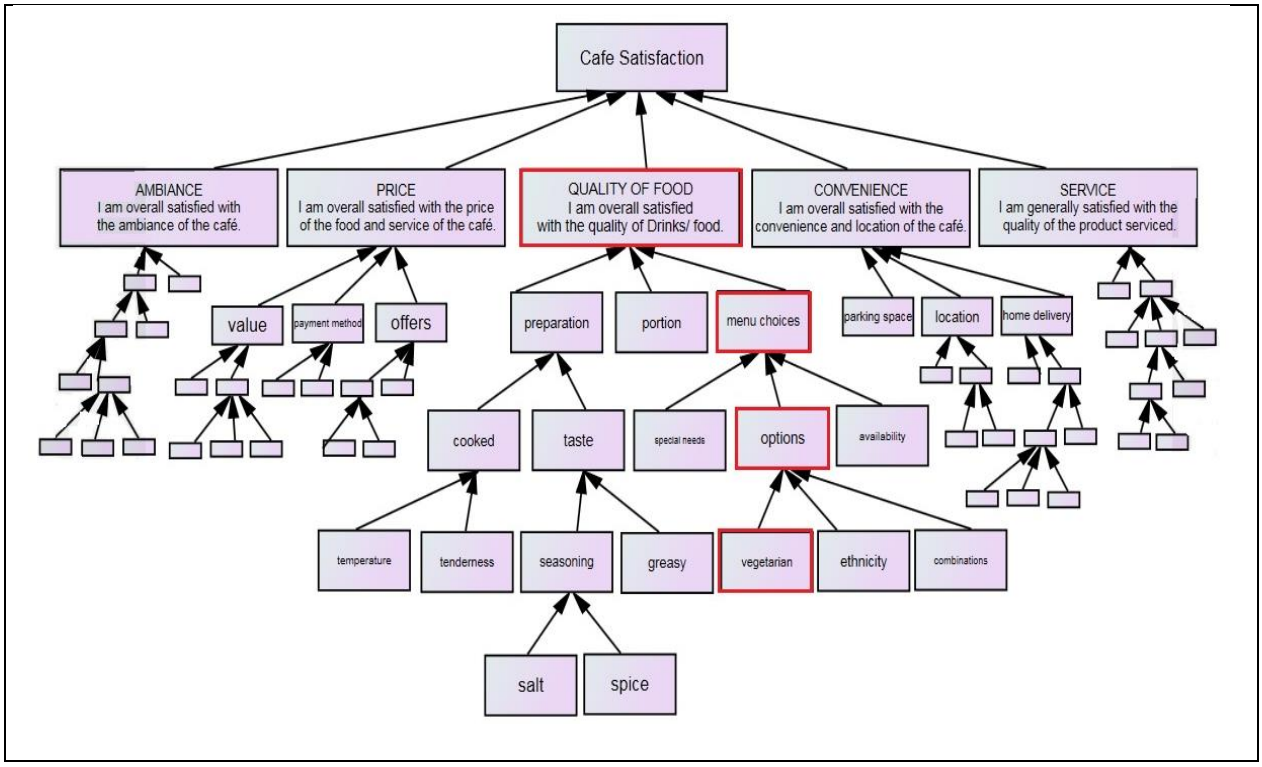


Figure 4.1. How CAS works for café satisfaction.

4.4 Assessing the Results of CAS

This paper solely focuses on assessing the conclusion validity of the results of CAS, which is approving the effectiveness of CAS as compared to psychometric surveys There are clearly other forms of validity, such as construct, internal, and external validity (Shadish et al., 2002). Testing these other forms of validity of CAS is the purview of our other research. Assessing the validity of CAS's results requires different techniques. As in CAS, there are child- parent relationships, where constructs that are children to other constructs in the tree should represent some dimension of the parent construct. As a result, in CAS, we can expect high correlations between a parent and one child (e.g., if a respondent says they are dissatisfied with service, there is at least one subdimension of service they are unhappy about). However, because the subdimensions are orthogonal, we expect low correlations between subdimensions (e.g., that someone is dissatisfied with the efficiency of service does not mean they are dissatisfied with the quality of service). Traditional statistical techniques cannot handle such complex correlations between items on a survey. As the goal of CAS is to measure individuals'

perceptions, CAS typically identifies one or a few narrowly defined constructs that respondents as a whole have the greatest or least affiliation to. Hence for CAS to have credible results would mean that the "correct construct(s)" have been identified. Thus, conclusion validity for CAS would mean that the results need to be checked and assessed against an external criterion which is a direct and independent measure of what the CAS is designed to measure.

One good example of an external criterion are online reviews. Online customer reviews are defined as "peer-generated product evaluations posted on company or third party web sites" (Mudambi & Schuff, 2010). Online reviews can both be very helpful, such as in the tourism context in which positive reviews can significantly increase the number of hotel bookings (Phillips, Zigan, Manuela, Silva, & Schegg, 2015), or in the context of online retail which demonstrates that sites with more helpful reviews offer greater potential value to customers (Mudambi & Schuff, 2010).

Online reviews can be used to assess a café satisfaction CAS for the following reasons. One, online customer reviews are considered as valuable sources of information for identifying relative strengths and weaknesses of products or service (Jindal & Liu, 2007; Somprasertsri & Lalitrojwong, 2010; Xu, Liao, Li, & Song, 2011) and they have been used for identifying the root cause of customer lack of satisfaction (Barreda & Bilgihan, 2013; Fu et al., 2013). Moreover, online reviews explicitly state what aspect of the product/service customers are not happy with. As an example, in one study, negative online consumer reviews were only considered because negative information is considered more diagnostic or informative than positive information (Lee, Park, & Han, 2008).

Two, online reviews both tend to have breadth and depth of information. Information depth and breadth refer to the extent to which information is sufficiently complete and exhaustive for a particular task (Wang & Strong, 1996). Depth of information in social media refers to the

richness of the content, and content breadth which is the number and diversity of perspectives. Both of these aspects can increase the diagnosticity of a review which provides a more accurate understanding of what aspect of the product or service is wrong (Mudambi & Schuff, 2010).

In this study, the results of a café satisfaction CAS will be assessed by (1) comparing the results of CAS to against a psychometric survey using the same item bank, and (2) comparing the results to online reviews.

Hypotheses 1-4 aim to compare the measurement properties of our café satisfaction CAS against a psychometric survey and hypothesis 5 aims to assess the results of our café satisfaction CAS with online reviews. We employ café satisfaction as a proxy context for evaluating the efficacy of CAS.

Property 1: Non-Response Rates. Low response rates are often associated with sample representativeness, which negatively influence statistical power and increase the size of confidence intervals. More importantly, low response rates can seriously influence the perceived credibility of studies' results (King & He, 2005). One supposed benefit of CAS is it will have a lower non-response rate than a psychometric survey, because respondents are required to answer fewer items to complete the survey. Therefore, we hypothesize:

H_1 Our café satisfaction CAS will have a higher response rate than the psychometric survey.

Property 2: Consistency in Response. Lack of consistent responses is an indication that items are being carelessly responded to, and is higher in longer surveys as the cognitive effort is greater (Krosnick, 1991). One of the advantages of CAS is to reduce the number of items a respondent is required to answer, which tends to reduce the cognitive effort. Hence, in a CAS about lack of satisfaction, a non-chosen item should mean the respondent was not dissatisfied with the parent of the item, and therefore the item itself. To test whether this is true, we compare the items unanswered by all respondents in CAS against responses in a psychometric survey.

A demonstration that the corresponding items in the psychometric survey tend towards "satisfaction" rather than "dissatisfaction" on the Likert scale would provide evidence CAS is performing as intended. Hence, we hypothesize:

H_2 For all non-chosen items in CAS, the corresponding item in the psychometric survey will tend towards "satisfaction" rather than "dissatisfaction."

Property 3: Item Discrimination. Response quality has been an important factor in survey research. As an example, the findings of one study indicated that most peoples' answers to survey questions are completely random and if the same people are asked the same question in repeated interviews, only about half give the same answer (Zaller & Feldman, 1992). In other studies, the findings indicated that in shorter surveys, respondents suffer less fatigue, hence there are less blank items and false responses (Deutskens et al., 2004; Galesic & Bosnjak, 2009).

Respondents may not produce quality responses for various reasons, such as lack of motivation or cognitive overload (Krosnick, 1991). Disengaged respondents may not provide quality responses; they may either leave items blank or provide false responses (Galesic & Bosnjak, 2009; Groves, 2006; Heerwegh & Loosveldt, 2006; Porter et al., 2004; Tourangeau, Couper, & Conrad, 2004). One example of false responses is answering a sequence of items in exactly the same way, i.e., straightlining (Fricker, Galesic, Tourangeau, & Yan, 2005). As an example, a psychometric survey with a 5-point Likert scale could have a straight line sequence of 5s, indicating the respondent filled those items without bothering to read the questions (Krosnick, 1991; Zhang and Conrad, 2013). Another example is choosing neutral response options (for example, "No opinion" and "Don't know" answer options) instead of substantive options such as only choosing the value 3 from a 5-point Likert scale. In addition, false responses could also be produced when respondents randomly choose an answer to avoid the substantial cognitive effort required of processing each questionnaire item (satisficing) (Krosnick 1991). Other

factors such as speeding (i.e., giving answers very quickly) (Zhang & Conrad, 2013), fatigue and boredom (Galesic & Bosnjak, 2009) accumulate throughout the survey, and decrease the willingness of respondents to invest in the effort needed for good quality answers.

CAS has two distinguishing features that manage such respondent error variance compared to psychometric surveys. One is the role of choice questions. When a respondent indicates equal dissatisfaction with two things, CAS prompts the respondent to identify the item they are least satisfied with. Second is items are administered so previous questions determine the next questions, hence increasing the amount of interaction with respondents and distinguishing the branches which require more attention. Studies have shown that interactions such as conditional branching can not only assist with decreasing nonresponse items, but also increase respondents' attention to each question and increasing the quality of responses (Krosnick 1991; Manfreda, Batagelj &Vehovar 2002).

In this study, one of the aims is to assess which instrument (CAS or psychometric survey) produces responses that better differentiate constructs associated with root cause. For this to be true, respondents should rate certain items more highly than others. We argue because respondents are more engaged with CAS, respondents will provide more truthful answers, and thus such discrimination between items will occur. Conversely, the length, and monotony of the psychometric survey causes the respondent to switch off. The respondent no longer answers questions in a psychometric survey in a focused way. Responses in a psychometric survey vary less, because the respondent is less engaged. Thus, we hypothesize:

H_3 In the café satisfaction CAS, the standard deviation within an individual respondent, across items will be higher than the psychometric survey.

Property 4: Item Response Variance. Similarly, when respondents are more engaged, they are able to focus more on the questions. This means we expect respondents will produce

answers in CAS that are "closer to the truth." By implication, if a café is genuinely poor at a particular area of satisfaction, we should receive more consistent feedback to that effect. Conversely, if a café is satisfactory in an area of satisfaction, CAS should more closely reflect that. In a psychometric survey, because the respondent is less engaged, there is more randomness in the answers, thereby increasing the error. Hence:

H_4 The café satisfaction CAS compared to the psychometric survey will have a lower standard deviation within items across respondents.

Property 5: Correspondence Between Response and Negative Opinion Sentences. In CAS, the frequency of responses to an item is supposed to reflect the general level of satisfaction or dissatisfaction with that item. Items that people do not respond to are those customers are satisfied with. If a customer satisfaction CAS is representative of customer dissatisfaction, then we would expect that the frequency of responses to an item would correspond to the frequency of that same item on online reviews. However, CAS is arranged in a tree. Some items in CAS are more granular than others. Similarly, reviews can be general, or specific. If a reviewer makes a specific comment, that comment should be able to be mapped to both the specific item in CAS, as well as the item's immediate parent and all further ancestors. However, if a reviewer makes a general comment, that comment will be unable to be mapped to specific comments in CAS. Therefore:

Proposition 1. There should be a good correspondence between the responses of subjects to CAS and online customer reviews.

H5: There is a high degree of correspondence between the top-level constructs in CAS and online reviews.

Let x refer to a level of the CAS hierarchy tree, where the depth of the hierarchy ranges from 2..n.

HAx: There is a high degree of correspondence between the xth level constructs and online reviews.

HBx: The degree of correspondence of the xth level constructs with online reviews will be lower than the degree of correspondence of the (x-1)th level.

We use the term "correspondence" here, because what we mean is the results of CAS should be "similar" to that of online reviews. However, the two may not be identical, as the population who perform online reviews are a specific subset of individuals who actually visit cafes. We are comparing two sets of frequencies from two samples (CAS and online reviews). If the two sets of frequencies are similar then they should have two characteristics, (1), the two samples should come from a common distribution and (2), there should be a high correlation between them. The first would indicate that there is an agreement with the constructs respondents are not happy with and the second is an indication of the level of agreement between the results of CAS and online reviews. As an example, if in the results of both café satisfaction CAS and online reviews, respondents agree that they are dissatisfied with the construct "price of coffee," then next step would be to assess the strength of their agreement. Thus, if the results of our CAS indicates that this construct (price of coffee) has been chosen the most and has an average score of 2.5, then one would expect many comments to be about the price of the coffee being too high.

4.5 Methodology

The remainder of this section describes how the sample, instrument, and data collection was conducted for each study.

4.5.1 Sample

Data was obtained from students at a public university in New Zealand. We selected 5 cafes in the university campus to compare and assess the results. We chose those cafes based on

which had the most available reviews on blog and review sites. We obtained online reviews for the 5 cafés from 4 popular customer review sites, Zomato, Yelp, Four Square and Trip Advisor (Sadhu, John, Kulkarni, & Tiwari, 2016). The target population was customers of the 5 cafes- i.e., university students. Our sample was students from the Information Systems and Operation Management (ISOM) Department and Economics Department. We were limited to only two departments due to conditions imposed by our ethics committee. There were approximately 5700 undergraduates and 120 postgraduates in both departments. An invitation to participate in our study was disseminated through the university student learning management system. As an incentive to participate, respondents were entered into a lucky draw worth 20 New Zealand dollars, and 20 respondents won the lucky draw.

4.5.2 Instrument

The item bank for the café satisfaction CAS and the psychometric survey consisted of 175 survey questions. It is key to note that thee psychometric survey the items were categorized similarly to the CAS and when respondents answered all the items from a category, they continued to the next set. The tools were developed as follows. First, we synthesized existing café satisfaction surveys (Hwang & Zhao, 2010; Kim et al., 2005; Liang & Zhang, 2009; Pizam & Ellis, 1999; Pratten, 2004; Ryu & (Shawn) Jang, 2008; Saglik et al., 2014; Shanka & Taylor, 2005). In addition, the first author trawled Internet café forums to identify common complaints. New items were developed based on those complaints. Here, principles from grounded theory (Strauss & Corbin, 1994) specifically, axial coding, guided us, as existing categories were saturated and items were explored and similarities were found. This continued until no more new conceptual categories emerged, i.e., theoretical saturation was achieved (Bowen, 2008). As an example, if we found in an internet café forum, customers were complaining about payment methods, we would create items about different possible ways of payment in cafés (e.g., cash, credit card, vouchers). This would continue until we had covered all existing ways. Approximately 400 items were collected. Items across the surveys and from

the forums were then compared and duplicates were discarded. Fewer than 300 items remained after this step.

Two independent raters blinded to the study's purpose went through the items and marked items that were either vague or repetitive. Approximately 60 items were dropped here. Next, we rearranged and reorganized the questions into a tree. Constructs are mapped together in a tree with constructs concerning higher level concepts linking to constructs with greater precision. Two blind and independent raters grouped and mapped the constructs together. The hierarchies were then transformed into a quantitative form, and that quantitative form was tested to determine the inter-rater reliability of the individual branches in the tree. The hierarchies were then successively transformed to test if they branched in the same way. We have assessed the instrument for construct and content validity and items which did not "fit in" the tree were dropped, leaving only 175 items.

4.5.3 Data collection

We collected our data from CAS and psychometric surveys in two waves from each instrument. We performed a non-response bias test across the two waves, with the second wave being a proxy for non-response. The test was performed by running an independent sample t-test for each item having at least 20 responses. For the café satisfaction CAS, we found the mean scores were not significantly different across all items and thus there was no evidence for non-response bias. However, for the psychometric survey, we found that for three items, the mean scores were significantly different. However, at a p-value of 0.05, we expect 1 in 20 tests to be significant. Out of 175 items, with each item having at least 20 responses, 3 item were significant, which is well within the margin of error of a test of significance (McHugh, 2011).

The data for the two studies was collected in the following manner. For the first study, respondents were first presented a consent form and agreed with it. Respondents were informed of the length of the survey (i.e., 175 items). We did this, because research suggests informing

subjects up front about this increases response rates (Marcus, Bosnjak, Lindner, Pilischenko, & Schutz, 2007). If the respondent discovers a survey is longer than anticipated, trust placed in the researcher by the respondent is revoked, leading to a higher nonresponse or an increase of dropouts (Heerwegh & Loosveldt, 2006). In addition, respondents were given an option of quitting the surveys at any time, and those who wished to drop out were given the option of providing an explanation for their decision.

Respondents then filled out some demographic details. They next chose one of the 5 mentioned cafés they wished to assess. Next, each respondent was assigned a random number (1 or 2) which determined which survey each respondent would be assigned to. The psychometric survey took approximately 20-40 minutes to finish and approximately 3 to 5 minutes was required for the CAS. We preformed two rounds of data collection. In the first round, the preliminary results from CAS indicated that "price and value" was the construct respondents were most dissatisfied with. However, the items for "price and value" were situated in the bottom of the psychometric survey, where there was a high item nonresponse rate. Hence, in the second round, we moved the items associated with "price and value" to the top of the psychometric survey and left CAS with no alterations. Note while this is a potential confound in our study, it improves the results of the psychometric survey and thus makes it more difficult for our hypotheses to be supported. Approximately 510 responses were collected from both instruments for all five cafés. In the first wave, 170 responses were collected for CAS and 142 responses for the psychometric survey. In the second, which was conducted 3 months after the first wave, 105 responses were collected for CAS and 93 responses were collected from the psychometric survey which concludes the second part.

After assessing the results, the overarching constructs "convenience" and "ambiance" did not have sufficient data for further assessment. As an example, in café 1 only one respondent chose the construct "convenience" and in café 4 only 4 respondents out of 46 respondents chose that

construct as the reason for their lack of satisfaction . Hence these constructs were dropped from the analysis and the constructs “service quality,” “food and drink quality,” and “price and value” for all cafés were selected for further assessment. It should be noted that the CAS results for the overarching constructs which were dropped had high mean scores which suggest respondents were not dissatisfied with either construct across the five cafes. Thus, dropping those constructs did not materially impact our analysis.

For the second study, reviews from review sites were transformed into “opinion sentences.” An opinion sentence “contains one or more product features and one or more opinion words” (Hu & Liu 2004, p. 172). First, we divided each online review into opinion sentences using the methodology of Hu and Liu, which is as follows. For each review, we identify a product/service feature and identify the associated opinion word(s) (such as an adjective). If there was enough context for it to have meaning on its own, then we accepted it as an opinion sentence. If there was not enough context than we combined two or more sentences and assessed comprehensibility. The average opinion sentence length was 14 words, which we deemed reasonable as a sentence length is normally between 8-20 words (Smith, 1961). We obtained a total of 441 opinion sentences.

Second, we employed two independent raters blinded to the purpose of the study to identify each opinion sentence as either negative or positive. Since the aim of the survey is to explore lack of satisfaction , positive opinion sentences were discarded. Rater 1 recorded 244 negative opinion sentences, while rater 2 recorded 235 negative opinion sentences. The inter-rater reliability was significant, and kappa was 0.833, above the recommended threshold of 0.7 (Landis & Koch, 1977). All negative opinion sentences that raters disagreed on were dropped, leaving 225 negative opinion sentences. Café 1-5 had a final total of 66, 67, 12, 67, and 15 negative opinion sentences respectively.

Third, the raters independently mapped each negative opinion sentence to every construct in the CAS. The raters first started with the five top level constructs (service, convenience, ambiance, food and drink quality, and price and value) and mapped each negative opinion sentence to one or more of the constructs, as seen in Table 4.1. We allowed raters to find no mappings, but there was no instance of such in our study. As an example, the comment "THE MOST OVER PRICED COFFEE I have ever had." was recorded in "Price and Value" by both raters. The ratings were assessed for inter-rated reliability. The level of agreement among the two independent raters for each top-level construct (service, convenience, ambiance, food and drink quality, and price and value) was calculated using Kappa, and had the following results, 0.820, 1, 0.747, 0.686, and 0.728. The agreement for the construct "convenience," was 1, because there were no negative ratings assigned by either rater to the construct. For the purposes of this study, the assignments by the two raters were analysed separately- no attempt was made to reconcile different categorizations by the raters.

Once the raters had completed the top-level mapping, they continued to map the lower level one level at a time. Each parent construct has at least two child constructs. For example, our ambiance construct has 38 children which are arranged in three levels, while "price and value" has 18 children of which three are located in the second level. Again, raters mapped every negative opinion sentence to every construct. Each mapping exercise was restricted to only the direct children of one construct. Thus, when raters mapped the children of ambiance, they did not consider the children for price. We encountered sample size problems doing this as in both CAS responses and online reviews, the number of respondents and online reviews per category decreases as we move to lower levels, fewer and fewer people respond to each question, because people are directed to other parts of the tree As a result, we could only fully assess the top-level constructs for all cafés, second level of "food and drinks quality" and only one branch of "price and value" for cafés 1,2, and 4.

4.5.4 Analysis

Property 1: Non-Response Rates. In our first hypothesis, we expected that our café satisfaction CAS will have a higher response rate than the psychometric survey. We have two data points- non-response for the CAS and non-response for the psychometric survey. Hence, statistical analysis to support the hypothesis is not possible. It should be noted that a comparative statistical analysis of non-response is generally impossible. However, we feel that just because a question cannot be answered numerically does not mean the question should not be asked. Instead, we qualitatively analysed comments given by respondents who had quit either instrument. As there were only comments for the psychometric survey, we went through the comments and grouped together comments which were similar. In addition, the first author's grouping was compared to the grouping of one independent rater blind to the study purpose to ensure inter-rater reliability. Kappa was 0.730, above the recommended threshold of 0.7 (Landis & Koch, 1977). We then gave each grouping a label to ascertain the causes for quitting.

Property 2: Consistency in Response. In our second hypothesis, we expect that for any item(s) in CAS that was not chosen by any respondent, the corresponding item in the psychometric survey will tend towards "satisfaction" rather than "lack of satisfaction ." To test this hypothesis, we calculated the mean and standard deviation of all items in the psychometric survey which had no corresponding score across all completed CAS. We then ran a one-tailed one-sample t-test against a mean of 3 (i.e., "average" on a 5-point Likert scale) to assess if the true mean of the sample is higher than the comparison value (3). We also counted the number of people who scored each item less than 3.

Property 3: Item Discrimination. In our third hypothesis, we expect the standard deviation within an individual, across items will be higher than in the psychometric survey. Hence in both instruments, we calculated the standard deviation of all items an individual filled across

the entire survey. We then calculated the mean and standard deviation of the standard deviations. Finally, a two sample one-tailed t-test was executed to compare the café satisfaction CAS and psychometric survey.

Property 4: Item Response Variance. Our fourth hypothesis is that the across-individual, within-item standard deviation in a café satisfaction CAS should be lower than the equivalent item in a psychometric survey. Thus, we first calculated the standard deviation within item across individuals for both instruments for all cafés. This produced two standard deviations per questionnaire item, one for CAS, one for the psychometric survey. It was not possible to perform a traditional parametric statistical analysis with this sample size. Instead, we performed a sign test (Dixon & Mood, 1946). We subtracted the standard deviation of the psychometric survey from CAS from each item and assessed the extent the signs were positive or negative. An overwhelming number of negative signs would suggest items in the psychometric survey tended towards a higher standard deviation.

Property 5: Correspondence Between Response and Negative Opinion Sentences. In our second study, we have (1) Likert scale scores of every item selected by respondents which are from 1 to 5, (2) frequency of responses to the items, and (3) frequency of opinion sentences that mapped to each item. We analysed the data in the following way as shown in Table 4.1. First, for each respondent, for every level, we gave the construct(s) with the lowest score a count of 1 and the rest a zero. This mimics the CAS process, which only is concerned with the construct a respondent is least satisfied with. As an example, if in one level a respondent gave four constructs the scores 2, 2, 4 and 5, then we would count the constructs as, 1, 1, 0 and 0. Second, for every level, for the child constructs of the same parents, we calculated the sum of the counts. We continued this for each construct.

As an example, for café one, for the top-level construct "price and value", rater 1 has 36 negative opinion sentences and rater 2 has 37 negative opinion sentences mapped. In CAS, there is 45 counts of the least scores. In total the construct "price and value" has the most mappings and the highest count of least scored in CAS which indicates "price and value" is the least satisfied construct.

Café	Construct	service	Convenience	ambiance	food & drinks	price &value
Café 1	CAS	5	1	16	20	45
	Rater 1	13	0	3	24	36
	Rater 2	10	0	3	18	37
Café 2	CAS	8	15	10	15	40
	Rater 1	7	0	9	28	35
	Rater 2	7	0	8	25	35
Café 3	CAS	8	7	33	13	39
	Rater 1	2	0	2	7	5
	Rater 2	1	0	2	7	4
Café 4	CAS	11	4	16	20	38
	Rater 1	15	1	12	26	16
	Rater 2	14	1	11	24	18
Café 5	CAS	5	5	19	10	22
	Rater 1	1	0	3	6	4
	Rater 2	1	0	3	6	6

Table 4.1. Frequency of Low Scores for CAS and Negative Opinion Sentences for Top-Level Constructs.

Third, we wanted to assess whether the frequency of responses to the items from the café satisfaction CAS for every level and the frequency of opinion sentences which mapped to each item come from a common distribution. Traditionally, the comparison of frequencies is done with goodness-of-fit tests. However, such tests examine whether two distributions are identical, not whether the distributions are "similar." Most goodness-of-fit tests (e.g., chi-square-based) do not work well with our data, because they require a large number of categories (Cheung & Rensvold, 2002). Our CAS has five top-level constructs, which produces a low degree of freedom (df=4). The number of subcategories for the lower levels range from two to nine. Thus, to test whether the two samples come from a common distribution, we employ a Kolmogorov-Smirnov two sample test (known as a D-statistic) (Massey 1951; Pettitt & Stephens 1977; Pratt & Gibbons 1981). The null hypothesis regarding the distributional form is rejected (i.e., the two distributions are not identical) if the D-statistic is greater than a critical

value. Thus, our fifth hypothesis is supported if the p-value is greater than 0.05 (i.e., the null hypothesis is not rejected). The two sample K-S test is ideal for our case, because it is robust to a smaller numbers of categories (even those as small as two) (Klotz, 1967; Pratt & Gibbons, 1981). Also, Kolmogorov-Smirnov two sample tests are robust to very small sample sizes (even as small as 10) (Birnbaum 1952; Massey 1951; Pettitt & Stephens 1977). We allow a Kolmogorov-Smirnov two sample test to be performed on the frequency of responses from top level constructs, second-level constructs of "food and drinks quality," and third-level constructs of "price and value" which are child constructs of the second-level parent construct "value," as these are the constructs with enough data. The second-level parent construct "value" was the only branch with mapped negative opinion sentences, hence the only branch with data. In addition, we recognize that failing to reject the null hypothesis is an unusual statistical approach. However, it is commonly employed for goodness-of-fit testing (Cochran, 1952).

Given that practices for goodness of fit testing deviate from traditional inferential statistics, we also attempted to assess the degree of correspondence between the two distributions with a more traditional measure. Hence, as a second test of correspondence, we performed Pearson Correlation tests between our constructs and the negative opinion sentences (Cohen 1988; Kline 1998). Pearson's r measures the strength and direction (decreasing or increasing, depending on the sign) of a linear relationship between two variables. A high, positive r would suggest the distributions are similar. These tests were only performed when the number of opinion sentences for each level of the hierarchy was at least 5x the number of constructs in the hierarchy tree. For example, if a parent construct had 3 children, then there had to be at least 15 opinion sentences mapped to the children for us to perform the analysis. 5x the number of cells is a commonly accepted guideline for the minimum sample size used for many statistical analyses (Bentler & Chou 1987; Straub & Gefen 2004; Thomas & Watson 2002).

It should be noted that a Pearson r test of the data suffers from its own limitations. Notably, a Pearson's r test is traditionally used to measure correspondence between two interval values. In our case, we are using it to compare two aggregated groups of frequencies. The key difference is that significance in the traditional Pearson r test is calculated based on the sample size. In our case, statistical significance is calculated based on the number of categories, and hence is independent of the sample size. This makes a calculation of statistical significance meaningless. Essentially, we recognize there are limitations in the statistical tests we employ for this study, but would highlight there do not appear to be better alternatives to answer our fundamental question.

4.6 Results

Property 1: Non-Response Rates. In our first study, for our first hypothesis, we expected that our café satisfaction CAS would have a higher response rate than the psychometric survey. For CAS, out of 275 respondents, 31 quit without providing a reason. For the psychometric survey of 235 respondents, 70 quit. 40 respondents quit the psychometric survey with a reason as tabulated in Table 4.2 and 30 respondents quit without providing a reason. As noted earlier, statistical analysis is not possible given the data, as there are only two data points. Nevertheless, our analysis of the qualitative responses given is telling. Notably, 60% of respondents in the psychometric survey quit the survey due to the large number of the items in the survey. The next most common reason given was only given by 7.5% of respondents. This suggests support for H1.

Reason for quitting	percentage	example
Number of items	60%	175 questions is too long
Long time	7.5%	The survey will take too long for me to do.
Repetitive	7.5%	too long, a lot of repetition, never ending
Have put enough effort	7.5%	I have already answered 50 ....
Not sufficient motive	7.5%	I am not gonna answer 175 questions
Boring	5%	I'm bored it's too long mate
Incentive not sufficient	5%	too many questions, not worth a voucher

Table 4.2. Presents the reasons for quitting the psychometric survey.

Property 2: Consistency in Response. In our second hypothesis, we expected that any item not chosen by any respondent doing CAS would have corresponding scores in a psychometric survey indicating that respondents were satisfied with the item. The number of items no respondent answered in the CAS for cafés 1 to 5, are respectively, 70, 48, 54, 47, and 71. Consider Figure 4.2 which presents the corresponding items to the ones no respondent answered in the CAS, for the psychometric survey for cafés 1-5. The red shades represent the responses with the value of 1, the purple shades responses with the value of 2, yellow 3, blue 4, and green 5 respectively. As illustrated, the majority of responses for those items are equivalent to three and above.

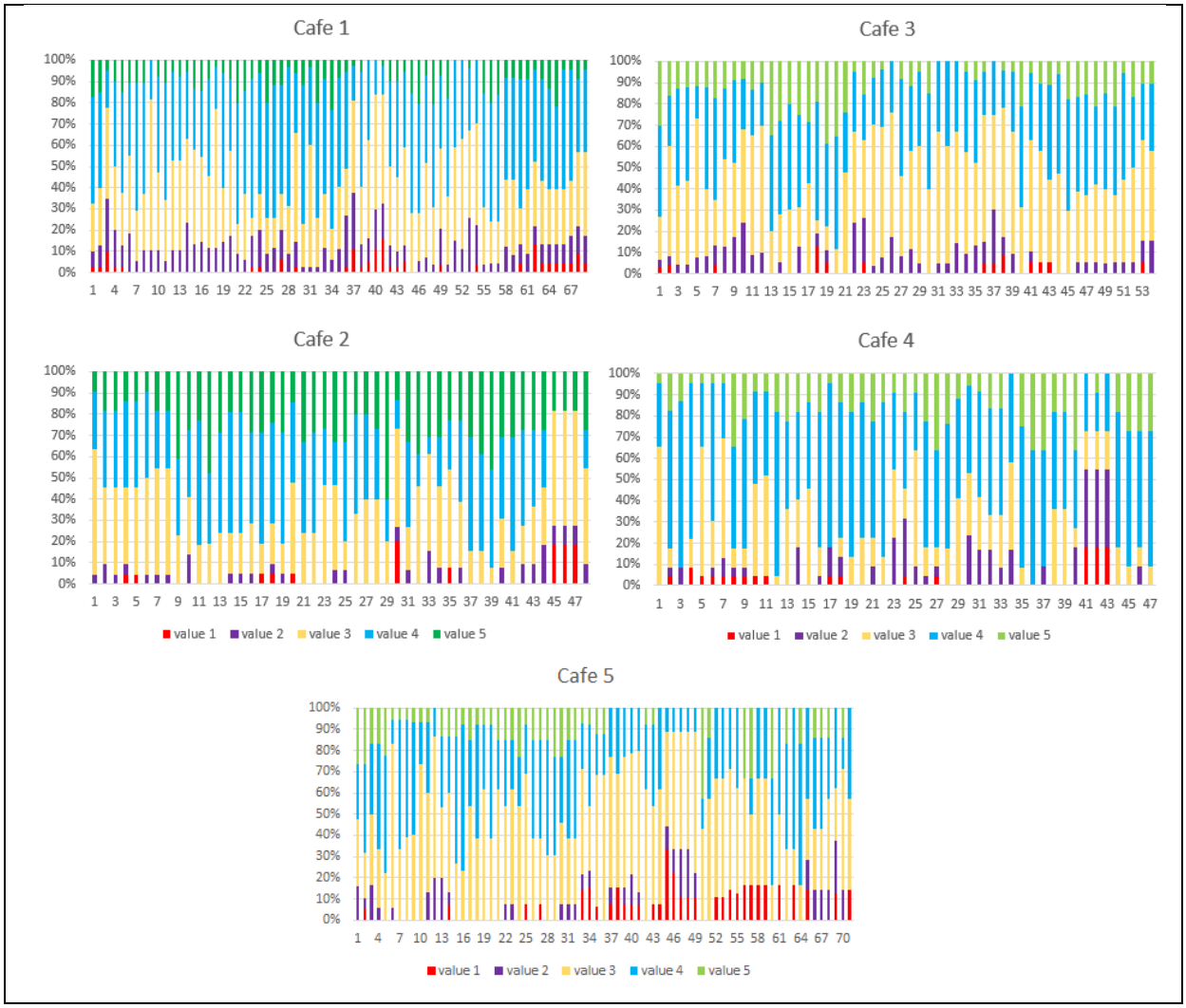


Figure 4.2. Presents the corresponding items no respondent answered in the café satisfaction CAS, for the psychometric survey for café 1-5.

As presented in Table 4.3, a one-tailed one-sample t-test against a mean of 3 (i.e., "average" on a 5-point Likert scale) of all the items which were null in CAS confirms this. In addition, we identified the number of items with a mean less than 3 for each café. About 5-11 percentage of items had a mean score of less than 3, the number dependent on the café analyzed. This indicates that the majority of the items not chosen in CAS were items that respondents were satisfied with; as those items in the psychometric survey were rated 3 and above on the Likert scale. Given the significance of the test results and relatively small percentage of responses indicating lack of satisfaction H2 is supported.

Cafe	Number of blank items in CAS	number of items with mean value < 3	Mean for psychometric survey Items	Std. Deviation	Test Value = 3		
					t	df	Sig. (2-tailed)
cafe1	70	4	3.4798	0.29189	13.15	69	<0.001
cafe2	48	4	3.7887	0.35517	15.385	47	<0.001
cafe3	54	0	3.5264	0.29615	13.061	53	<0.001
cafe4	47	3	3.7028	0.41698	11.555	46	<0.001
cafe5	71	8	3.3646	0.40286	7.625	70	<0.001

Table 4.3. Presents the t-test of all the items which were non-chosen in CAS for all psychometric surveys for the five cafés.

Property 3: Item Discrimination. In our third hypothesis, we expected the standard deviation across all items answered by an individual will be higher in CAS than the psychometric survey. Figure 4.3 presents the responses for both instruments for café 1. The grey shades represent the responses which are either item non-response in the psychometric survey or have not been chosen by respondents in the café satisfaction CAS. The red shades represent the responses with the value of 1, the purple shades responses with the value of 2, yellow 3, blue 4, and green 5 respectively. Cafés 2-5 share a similar pattern with café 1. As the Figure demonstrates, only a fraction of the items is responded to in CAS. The construct "price and value" is indicated as the one respondents are least satisfied with in both instruments as it has been answered by most respondents. In addition, in the café satisfaction CAS, certain items of the construct "price and value" show larger patches of red and purple compared to the psychometric survey, which means respondents were very unhappy with these items. The graph suggests CAS will have a higher cross-item standard deviation.

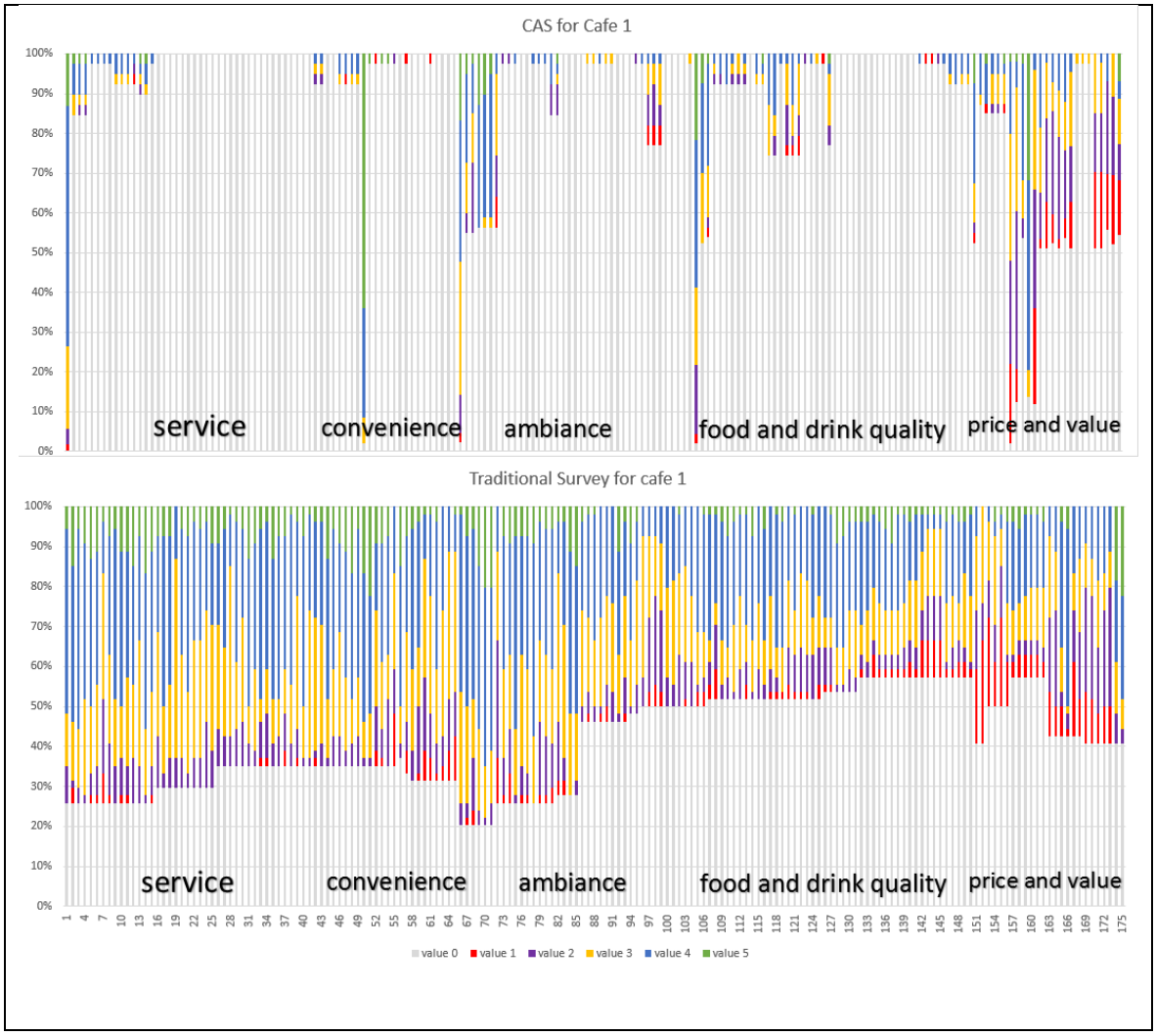


Figure 4.3. Presents the responses for both instruments for café 1.

We ran a two sample one-tail t-test of the mean standard deviation of every individual for all five cafes. According to Table 4.4, throughout the overarching constructs “service quality”, “food and drinks quality” and “price and value” café satisfaction CAS has a higher standard deviation across items. Hence H3 is supported.

Construct	**Overarching Constructs**	
Type	*CAS*	Psychometric survey
Mean	1.0113	0.656
Standard dev.	0.083	0.042
Observations	15 (3x5)	
Hypothesized Mean Difference	0	
t Critical one-tail	1.7081	
***p* (one-tail)**	0.0003	

Table 4.4. Compares the standard deviation of individuals across items of CAS with the items of the psychometric survey for the 5 cafes.

Property 4: Item Response Variance In our fourth hypothesis, we expected the café satisfaction CAS would have a lower within item standard deviation than the psychometric survey, which reflects the error the psychometric survey encourages. Table 4.5 presents the results of the sign test for the standard deviation of each survey question (175 items) for both café satisfaction CAS and the psychometric survey. As the number of negative sign counts are overwhelmingly higher for each cafe, H4 is clearly supported.

Café		**Positive sign count**	**Negative sign count**	***p*-value**
Café 1	CAS	11	164	<0.0001
	Psychometric Survey			
Café 2	CAS	8	167	<0.0001
	Psychometric Survey			
Café 3	CAS	4	174	<0.0001
	Psychometric Survey			
Café 4	CAS	2	173	<0.0001
	Psychometric Survey			
Café 5	CAS	13	162	<0.0001
	Psychometric Survey			

Table 4.5. Sign Test for the standard deviation of each item of CAS with each item of the psychometric survey for the 5 cafes.

Property 5: Correspondence Between Response and Negative Opinion Sentences. For our second study, in regards to our fifth hypothesis for top level parent constructs, we expected that our café satisfaction CAS would have a high degree of correspondence. As demonstrated in Table 4.6, the D-statistic is lower than the D-critical for all 5 cafes, and hence the p-values are

all above 0.05. Therefore, the sample from CAS and the sample from the online reviews appear to come from a common distribution.

Café	Rater	Alpha	D-statistic	D-critical	*p* for K-S	Sample size for CAS	Sample size for Reviews	r	df	*p* for r
café 1	R1	0.05	0.139	0.217	0.430	87	66	0.862	4	0.060
	R2	0.05	0.121	0.227	0.672	87	56	0.919	4	0.0275
Café 2	R 1	0.05	0.155	0.216	0.297	88	66	0.919	4	0.0275
	R 2	0.05	0.149	0.220	0.371	88	62	0.734	4	0.792
Café 3	R1	0.05	0.172	0.354	0.775	90	16	---	-	---
	R2	0.05	0.208	0.374	0.618	90	14	---	-	---
Café 4	R 1	0.05	0.198	0.212	0.079	89	70	0.53	4	0.358
	R 2	0.05	0.162	0.214	0.239	89	68	0.657	4	0.228
café 5	R 1	0.05	0.189	0.385	0.762	61	14	---	-	---
	R2	0.05	0.225	0.366	0.486	61	16	---	-	---

Table 4.6. Degree of Correspondence for the top-level constructs.

Our Pearson correlation tests produce similar results as demonstrated in Table 4.6. Cafés 3 and 5 each had less than 25 negative opinion statements (5x5) mapped to the top-level constructs, hence they had an insufficient sample size and were omitted. Observe how in the table, the Pearson r produces very high correlations (all r> 0.6). Cohen (1988) notes that an r >0.5 is a strong correlation. This suggest support for our hypothesis of correspondence between top level constructs and online reviews for cafes 1, 2 and 4. Note the statistic r is a measure of effect size and is sample-size neutral. The statistic *p* for the Pearson r is not significant, because the *p*-value is calculated based on the number of categories, not on the sample size. The degrees of freedom is 4, because there are 5 categories. The Pearson r test of significance does not take into account that we really had 53 respondents for each café.

In regards to our 5th hypothesis for lower level parent constructs, we expected that our café satisfaction CAS would have a lower degree of correspondence than the higher level parent constructs. For the overarching construct “food and drinks quality” as seen in Table 4.7, for cafes, 2 and 4 the K-S *p* value results suggest almost all the CAS constructs and online reviews

come from the same distribution except for rater 1 in café 1. While, for cafés 2 and 4, results suggest a high correlation between CAS and online reviews, for café 1 Pearson r is just moderate. This suggests that the construct "food and drinks quality" and its attendant subconstructs capture the essence of dissatisfaction associated with food and drink quality.

For the overarching construct "price and value," a K-S two sample test was only suitable for one of the three second-level constructs, as the other constructs (i.e., value) did not have sufficient data from online reviews. We performed a K-S two sample test and Pearson Correlation on the subconstructs of value. As Table 4.7 demonstrates, the results are mixed for café 1, not supported for café 2, but supported for café 4.

Café	**Construct**	**Rater**	**Sample size for CAS**	**Sample size for Reviews**	**D-statistical**	**D-critical**	***p* for K-S**	**r**	**df**
Café 1	" food & drinks quality"	R 1	22	20	0.423	0.4	0.033	0.502	2
		R 2	20	20	0.4	0.409	0.059	0.458	2
Café 2	"food & drinks quality"	R 1	15	26	0.259	0.419	0.508	0.769	2
		R 2	15	24	0.267	0.425	0.462	0.712	2
Café 4	"food & drinks quality"	R 1	20	34	0.121	0.367	0.988	0.923	2
		R 2	20	31	0.102	0.373	0.999	0.953	2
Café 1	"value"	R 1	87	40	0.257	0.26	0.054	0.457	2
		R 2	87	29	0.494	0.282	0.001	0.526	2
Café 2	"value"	R 1	84	38	0.26	0.258	0.040	0.324	2
		R 2	84	26	0.5	0.3	0.001	0.363	2
Café 4	"value"	R 1	71	26	0.263	0.3019	0.119	0.662	2
		R 2	71	26	0.302	0.303	0.05	0.694	2

Table 4.7. Degree of Correspondence for "food and drinks quality" and "value".

4.7 Discussion and Conclusion

In our first study, we compared a café satisfaction CAS to a psychometric survey of the same item bank. As our study demonstrates, CAS has certain advantages over the psychometric survey. First, given the same item bank, CAS tends to have a higher response rate as respondents are required to respond to fewer items (Galesic & Bosnjak, 2009). Second, the

CAS instrument can function as intended to identify constructs which respondents have the greatest or least affiliation to. When respondents do not fill out CAS items, it really means those items are unimportant for further analysis. Third, the cross-item standard deviation in CAS is higher than in a psychometric survey, which means it is easier to identify salient items in CAS than in a psychometric survey. Finally, CAS has a much lower random error rate, possibly because of reduced respondent fatigue, than a psychometric survey. The CAS within-item standard deviation is much lower than in a psychometric survey.

In our second study, our results demonstrate that our top-level constructs from CAS have a high degree of correspondence with online reviews. However, we could not adequately validate the lower level constructs due to a lack of negative opinion sentences that mapped to those constructs. There are several reasons for this. First, some online reviews raised problems only at the level of the top -level constructs. For example, a reviewer might say, "I really don't like their coffee or any of their desserts." Notice how the reviewer complains about the poor quality of the food and drink but does not break down why he or she is dissatisfied Second, some online reviews only fully cover a single complaint about a café in depth. As an example, a reviewer might say "The cheesecake tasted great the first few bites but towards the end the layer of milk chocolate and the rich cheesecake becomes overwhelmingly sweet and rich." In this quote, the reviewer has only complained in detail about the cheesecake, and nothing else. It is possible the reviewer is dissatisfied with other elements of the café but chooses not to talk about it.

Why is this? Most online reviews are posted on the company or third-party website and they have certain limitations. One, these environments often limit the number of words in the review. As an example the average word count of Amazon reviews is 181.5 words (Wu & Huberman, 2008). That word count impacts review quality is well documented, as longer reviews generally increase the helpfulness of the review (Mudambi & Schuff, 2010).

According to Salehan and Kim (2016), longer reviews are expected to contain more information and thus attract more readerships. In addition, longer reviews are more likely to analyse different aspects of the product which leads to increased perceptions regarding their helpfulness (Chevalier & Mayzlin, 2006; Salehan & Kim, 2016). As an example, the usefulness of reviews are expected to increase by 19.9% when the reviewers write 83 words more than the average length of reviews (135 words) (López & Farzan, 2014). Simply put, the average online review is of low quality. Two, in online reviews, the number of positive statements is greater than negative statements. For instance, in one study, reviews were divided into component statements, which are "defined as group of words that comprise a single thought"(Schindler & Bickart, 2012). According to Schindler and Bickart (2012) there are five statement types in online reviews, (1) Positive evaluative statements, (2) Negative evaluative statements, (3) Product-descriptive statements, (4) Reviewer-descriptive statements, and (5) Other statements. Their study indicated that there were more positive evaluative statements than negative evaluative statements which indicates that there are often more positive reviews than negative ones, as people, in general, are more likely to provide online reviews when they are generally satisfied (Godes & Mayzlin, 2004; Schindler & Bickart, 2012).

This suggests that CAS is actually better than an analysis of online reviews for diagnosing problems with café satisfaction. Results of this study suggests that CAS can be applied as a useful instrument to further a number of areas of research. Psychometric survey research has been useful for developing causal relationships between constructs, but has been poor at "unpacking" constructs to develop a rich understanding of how things work. As an example, Bagozzi (2007) highlights that research on the technology acceptance model has demonstrated the relationship between perceived usefulness, ease of use, and intention to use, but cannot articulate why this relationship holds. CAS provides a quantitative tool for doing this. Perceived usefulness and perceived ease of use can be redefined as CAS constructs to identify

why people do not perceive a piece of IT as useful or easy to use. Just as structural equation modelling enabled researchers to study complex causal structures, CAS enables researchers to begin quantitatively exploring questions of why, thereby furthering theory (Sutton & Staw, 1995).

As future research, we hope to explore and assess CAS in several areas. One, as we are not able to fully assess the conclusion validity of CAS, future work is required to develop techniques for assessing the credibility of CAS results. Two, we want to assess CAS in other fields. This study compared a café satisfaction Computer-Adaptive Surveys (CAS) against a psychometric survey of the same item bank. We chose café satisfaction because of the availability of existing, relevant instruments. If we were to develop our own instrument, the potential quality of the questionnaire items would be a confound. By using existing items employed in psychometric surveys, this confound is eliminated. Future work will apply CAS to develop CAS in other domains such as employee satisfaction or in health sectors. Three, we continue developing techniques, strategies, and algorithms to increase the accuracy and efficiency of CAS. This study used an adaptive version of branching for respondents to move from one set of items to another set according to a pre-defined criterion. Future study aims to assess and explore other possible options for determining the next set of question(s).

References

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, 8(4), 244–254.

Barreda, A., & Bilgihan, A. (2013). An analysis of user-generated content for hotel experiences. *Journal of Hospitality and Tourism Technology*, *4*(3), 263–280.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117.

Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association*, *47*(259), 425–441. Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research*, *8*(1), 137–152.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*,

9(2), 233–255.
Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.
Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, *23*(3), 315–345.
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*(1), 21–36.
Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, *41*(236), 557–566.Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *The Public Opinion Quarterly*, *69*(3), 370–392.
Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., & Sadeh, N. (2013). Why people hate your app. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, 1276.
Galesic, M., & 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.
Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*(4), 545–560.
Goodman, J. A., Broetzmann, S. M., & Adamson, C. (1992). Ineffective - That's the problem with customer satisfaction surveys. *Quality Progress*, *25*(5), 35–38.
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., & 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. 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.
Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 04*, *4*, 168.
Hwang, J., & Zhao, J. (2010). Factors influencing customer satisfaction or dissatisfaction in restaurtant business using answertree methodology. *Journal of Quality Assurance in Hostpitality & Tourism*, *11*(2), 93–110.
Jindal, N., & Liu, B. (2007). Analyzing and detecting review spam. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 547–552.
Kim, Ay.-S., Moreo, P. J., & Yeh, R. J. M. (2005). Customers' satisfaction factors regarding university food court service. *Journal of Foodservice Business Research*, *7*(4), 97–110.
King, W. R., & He, J. (2005). External validity in is survey research. *Communications of the AIS*, *16*, 880–894.
Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling* (Third). New York: The Guilford Press.
Klotz, J. (1967). Asymptotic efficiency of the two sample Kolmogorov-Smirnov test. *Journal of the American Statistical Association*, *62*(319), 932–938.
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude

measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
Lee, J., Park, D., & Han, I. (2008). The effect of negative online consumer reviews on product attitude : An information processing view. *Electronic Commerce Research and Applications*, *7*(3), 341–352.
Liang, X., & Zhang, S. (2009). Investigation of customer satisfaction in student food service. *International Journal of Quality and Service Sciences*, *1*(1), 113–124.
López, C., & Farzan, R. (2014). Analysis of local online review systems as digital word-of-mouth. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 457–462).
Manfreda, K. L. M., Batagelj, Z., & Vehovar, V. (2002). Design of websurvey questionnaires: three basic experiments. *Jounal of Computer-Mediated Communication*, *7*(3), 0.
Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schutz, a. (2007). Compensating for low topic interest and long surveys: a field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*, 372–383.
Massey, F. J. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.
McHugh, M. L. (2011). Multiple comparison analysis testing in ANOVA. *Biochemia Medica*, *21*(3), 203–9.
Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185–200.
Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, *20*(1), 61–71.
Pettitt, A. N., & Stephens, M. A. (1977). The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, *19*(2), 205–210.
Phillips, P., Zigan, K., Manuela, M., Silva, S., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management*, *50*, 2010–2012.
Pizam, A., & Ellis, T. (1999). Customer satisfaction and its measurement in hospitality enterprises. *International Journal of Contemporary Hospitality Management*, *11*(7), 326–339.
Pinsonneault, A., & Kraemer, K. (1993). *S*urvey research methodology in management information systems: an assessment. *Journal of Management Information Systems, 1*0(2), 75-105.
Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, *2004*(121), 63–73.
Pratt, J. W., & Gibbons, J. D. (1981). Kolmogorov-Smirnov two-sample tests. In *Concepts of Nonparametric Theory* (pp. 318–344). New York: Springer.
Pratten, J. D. (2004). Customer satisfaction and waiting staff. *International Journal of Contemporary Hospitality Management*, *16*(6), 385–388.
Rooney, J. J., & Van den Heuvel, L. N. (2004). Root Cause Analysis for Beginners. *Quality Progress*, (July), 45–53.
Ryu, K., & (Shawn) Jang, S. (2008). DINESCAPE: a scale for customers' perception of dining environments. *Journal of Foodservice Business Research*, *11*(1), 2–22.
Sadhu, D., John, S., Kulkarni, R., & Tiwari, A. (2016). Mining and predicting reviews to micro-reviews and detection of manipulated reviews for hotel websites. *International Journal of Emerging Technology and Computer Science*, *1*(2), 28–31.
Saglik, E., Gulluce, A. C., Kaya, U., & Ozhan, K. C. (2014). Service quality and customer

satisfaction relationship : A research in Erzurum Ataturk University. *American International Journal of Contemporary Research*, *4*(1), 100–117.
Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, *81*, 30–40.
Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234–243.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental for Generalized Designs Causal Inference*. *Wadsworth Cengage learning*. Houghton Mifflin Company.
Shanka, T., & Taylor, R. (2005). Assessment of university campus café service: The students' perceptions. *Asia Pacific Journal of Tourism Research*, *10*(March 2015), 329–340.
Smith, E. A. (1961). Devereux readability index. *The Journal of Educational Research*, *54*(8), 298–303.
Somprasertsri, G., & Lalitrojwong, P. (2010). Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal Computer Science*, *16*(6), 938–955.
Straub, D., & Gefen, D. (2004). Validation guidelines for is positivist research. *Communications of the Association for Information Systems*, *13*, 380–427.
Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of Qualitative Research*, 273–285.
Sutton, R. I., & Staw, B. M. (1995). What theory is not. *Administrative Science Quarterly*, *40*(3), 371–384.
Thomas, D. M., & Watson, R. T. (2002). Q-Sorting and Mis research: A primer. *Communications of the Association for Information Systems*, *8*, 141–156.
Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing , position , and order interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393.
Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Source Journal of Management Information Systems*, *12*(4), 5–33.
Wu, F., & Huberman, B. A. (2008). How public opinion forms. *In International Workshop on Internet and Network Economics* (pp. 334–341).
Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems*, *50*(4), 743–754.
Zaller, J., & Feldman, S. (1992). A simple theory of the survey response : Answering questions versus revealing preferences. *American Journal of Political Science*, *36*(3), 579–616.
Zhang, C., & Conrad, F. G. (2013). Speeding in web surveys : The tendency to answer very fast and its association with straightlining. *European Survey Research Association*, *8*(2), 127–135.

5 Conclusion

Computer-Adaptive Surveys are new instruments that can potentially represent and test diagnostic theories. However, validation methods for CAS are under researched and non-existent. This study had three aims, one, develop a method to compare two CAS trees , two create a CAS tree by using a variant q-sort technique which is designed to assess the validity of CAS trees, and finally assess the ability and efficacy of CAS. Each aim is mapped to a chapter of the thesis, through chapters two to four, I demonstrated that Computer-Adaptive Surveys (CAS) can make a great contribution to research methodology. Specifically:

Chapter 2 presented a boot strapping technique to demonstrate how three measures Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma change as a hypothetical CAS tree deviates from a "true" version. I created two general types of modifications, movements and swaps. I repeated those modifications many times and implemented this on various "perfect" trees of various sizes.

I found that each measure provided certain information. One, Lambda is useful for determining two things, one whether the principle problem is disagreement among single nodes (type 2 movements), which indicates that while raters agree on grouping of child nodes, they disagree on the parent of the child nodes. Two, a high Lambda in concurrent with a low Kappa or Gamma is useful to detect swaps. Two, Gamma is useful for identifying disagreements with the hierarchy of the constructs. Finally, Kappa is useful for identifying disagreements on the mapping of constructs of the same level.

I then proposed thresholds for various levels of inter-rater reliability, as an example, a threshold for when Lambda > 0.7, Kappa > 0.4 and Gamma > 0.3, suggest there is no more than 30 percent of modification between two CAS trees.

Chapter 3 presented a step by step variant q-sort technique designed to assess the validity of CAS trees. In my variant q-sort technique, trees that independent and blind raters developed were transformed into a table and then, the similarity of the trees was measured by using three measures of association, Goodman and Kruskal's Lambda, Cohen's Kappa, and Goodman and Kruskal's Gamma (Goodman & Kruskal, 1963). These measures together provided useful information for evaluating the similarity of two tress and also diagnosing the causes of the differences.

Chapter 4 presented two studies to assess the credibility of CAS results. In the first study, I compared a café satisfaction CAS to a psychometric survey of the same item bank. As the thesis demonstrates, CAS has certain advantages over the psychometric survey. One, CAS tends to have a higher response rate as respondents are required to respond to fewer items (Galesic & Bosnjak, 2009). Two, the CAS can function as intended to identify constructs which respondents have the greatest or least affiliation to. When respondents do not fill out CAS items, it really means those items are unimportant for further analysis. Third, the cross-item standard deviation in CAS is higher than in a psychometric survey, which means it is easier to identify relevant items in CAS than in a psychometric survey. Finally, CAS has a much lower random error rate, possibly because of reduced respondent fatigue, than a psychometric survey. The CAS within-item standard deviation is much lower than in a psychometric survey. My second study in chapter 4, demonstrated that my top-level constructs from CAS have a high degree of correspondence with online reviews. However, I could not adequately validate the lower level constructs due to a lack of negative opinion sentences that mapped to those constructs.

In my thesis, I demonstrated that CAS is superior to a traditional long survey. It could be argued that I attacked a strawman, because few surveys of such length are actually administered, such as the work of Galesic and Bosnjak (2009) who implemented 180

questions divided into 20 blocks. The traditional alternative strategy using a traditional survey would be to issue what would be equivalent to a top-level CAS to one small group of respondents, and then the corresponding second-level CAS to a second wave and so on. However, this approach is inferior to CAS, because it necessitates dividing up one's sample. The actual sample size performing the survey is thus a fraction of CAS.

Consider as an example a CAS with 5 levels on a sample size of 100. When one administers the CAS, all 100 people traverse all 5 levels. In contrast, with a traditional survey, 20 people would take the top-level survey, another 20 would take the second level etc. It could be that the 20 people taking the second level survey do not agree with the first 20 people on what is the most severe issue.

In my thesis, I also demonstrated that CAS is superior to online ratings. Online customer reviews are considered as valuable sources of information for identifying relative strengths and weaknesses of products or service (Jindal & Liu, 2007; Somprasertsri & Lalitrojwong, 2010; Xu, Liao, Li, & Song, 2011) and they have been used for identifying the root cause of customer dissatisfaction (Barreda & Bilgihan, 2013; Fu et al., 2013). Moreover, online reviews explicitly state what aspect of the product/service customers are not happy with. However, CAS is a better tool for diagnosing problems than online reviews for several reasons. One, CAS in our study provided a broader breadth and richer depth of information compared to online reviews. Two, CAS can be useful for times where there is not a lot of online reviews, as methods such text mining require a large sample size. Finally, in CAS, the aggregate of the results identify the most pressing issues in a quantitative form.

There are, of course, other potential competing methodologies such as focus groups or interviews. CAS is superior to focus groups and interviews in terms of cost. As demonstrated in my thesis, CAS allows one to assess cause at a level of depth superior to online

ratings. While it is possible to obtain deep insight from focus groups or interviews (Myers & Newman, 2007), it is normally expensive to perform such. Most qualitative research such as grounded theory methodology is able to interview at most 50 participants (Morse, 1994, p.225). In contrast, in my thesis alone, I sampled 510 people

5.1 Limitations

This section summarizes the limitations of the research and CAS which were presented in the individual articles.

5.1.1 Limitations of study

This study has several limitations. One, there are many forms of validity, such as construct, internal, and external validity (Shadish, Cook, & Campbell, 2002), which this thesis did not assess. This study solely focuses on assessing the validity of the CAS tree and results of CAS.

Two, in CAS respondents are not required to respond to all of the items in the CAS tree. Findings of other studies have indicated that in shorter surveys, respondents suffer less fatigue, hence there are less blank items and false responses (Deutskens, de Ruyter, Wetzels, & Oosterveld, 2004; Galesic & Bosnjak, 2009). However, this study does not assess or measure the amount of fatigue in the respondents in both CAS and the psychometric survey.

Three, in this study, the focus was to use a CAS focused on identifying the root cause of customers' lack of satisfaction with a café and CAS did not assess the constructs respondents were satisfied with.

Finally, in this study, the algorithm used in CAS was designed to identify only two out of five top-level constructs with the lowest scores. This means that the top-level construct with the least score was identified first, and the respondent responded to the items relevant to that top-

level construct. Then the next least scored top-level construct was identified and the related items would be responded to. Hence respondents would go through CAS only twice.

5.1.2 Limitations of CAS

CAS has several limitations. Firstly, CAS has a large item bank which normally consists of hundreds of items. A key limitation in CAS is that respondents only answer the items that are most related to the dimensions in which the respondent has the most extreme perceptions about. Hence, not all items are responded to. This does not imply that the other items are irrelevant, but one or more items are more relevant and thus scored more extreme. The key assumption of CAS is that the items with the most extreme scores are identified so that we can target the principal problem first, as fixing the principal problem often times can resolve other issues.

Secondly, CAS has a very specific narrow role (the limitation). In a CAS, there is an implicit independent variable (e.g., customer lack of satisfaction) that the survey does not ask. It is used to identify root cause as CAS attempts to identify which construct has the most exploratory power in identifying the root cause of that dependent variable, i.e., why are customers unhappy with the café? Unlike other surveys which assess the correlations among constructs so that a model can be either assessed or created.

Thirdly, many nomological properties of the CAS trees remain unknown Nomological validity refers to the degree to which the constructs fit within the logical network of the theory is nomological validity (Calder, Phillips, & Tybout, 1983; Mackenzie, Podsakoff, & Podsakoff, 2011). In other words, it is a measure of the theoretical correspondence between the theory and the constructs within the theory. However, there are a very few statistical tests of nomological validity for formative items (Bollen, 1989; A Diamantopoulos & Winklhofer, 2001; Adamantios Diamantopoulos, Riefler, & Roth, 2008; Freeze & Raschke, 2007). In addition,

assessing the nomological validity of CAS Tree requires different techniques as items are mapped in a tree and consist of hundreds of items.

Fourthly, in a CAS Tree, the number of children per parent is limited. As the thresholds would not be appropriate if there were more than 7 children per parent.

Finally, in contrast to psychometric surveys which only require a pen and paper, in CAS, respondents are required to have a smart device (such as a mobile or laptop) in order to access CAS and respond to the questions.

5.2 CAS in Information Systems (IS)

CAS can be used in various IS context. I present two examples, the first is how to develop a CAS instrument to identify the principal root cause for why Kanban is unsuccessful in organizations and the second example is how to build a CAS tree on self-efficacy for Instagram.

5.2.1 Example 1

Flow techniques are gaining popularity across the ISD community, as managing a continuous and smooth flow of value creating activities that deliver value to the customer is a core principle of lean (Anderson et al. 2011). There are several examples which improve flow in the software development process such as Kanban. One issue unaddressed in IS research is why Kanban adoption has been difficult. Kanban is useful for many reasons, such as to improving team communication, development flow, reducing time to reach the market, increasing productivity, and creating transparency in organization (Ahmad et al. 2014).

However, there have been many reported challenges with Kanban in organizations (Ahmad et al., 2018; Ahmad, Markkula, & Oivo, 2013; Ahmad, Markkula, Oivo, & Kuvaja, 2014; Balve, Krüger, & Tolstrup Sørensen, 2017; Dennehy & Conboy, 2016, 2017; Tanner & Dauane, 2017).

Many reasons have been given for this failure, for instance, according to Dennehy and Conboy

(2016), in their study , one user complained that the Kanban board " looked very gimmicky and I just don't trust pieces of paper stuck to the wall". In each organization, individuals may have different reasons for not willing to use that particular piece of technology. In these cases, as researchers, we would want to know why and explore those reasons.

There are many theories such as Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al. 2003), Technology Culture Conflict (Leidner & Kayworth, 2006), Self-Efficacy (Scott & Walczak 2008), Technochange Management (Markus, 2004), and Virtual Team Management (Maruping et al., 2009; Wiener et al., 2016) that have done substantial research to explain why such problems exist. As an example, self-efficacy is unpacked to prior experience, computer anxiety, organizational support, and engagement; where the challenge, "lack of experience with the Kanban method" may be mapped to prior experience and "lack of support from the management" may be mapped to organizational support. While there has been substantial work on developing various theories, little research has attempted to unpack these challenges to identify the root cause of why such challenges exist (Bagozzi, 2007; Legris, Ingham, & Collerette, 2003). There are numerous studies that test a theory(s) such as the work of Dennehy and Conboy (2017) in Activity Theory or Delone and McLean IS success model. However, in this case, the question is not about which theory is right, but as which theory is most relevant in identifying the root cause. Hence this study takes a multiple theory testing approach to assess which theory can best identify the root cause of why Kanban is unsuccessful in organizations.

To do this, I first develop a framework, in which I will be employing a CAS tree which has been pioneered for creating a hierarchy for mapping constructs. The hierarchies are built in the following way, first, each top-level construct represents a theory and should capture the entire scope of the problem domain, such as UTAUT for Kanban. The notion of scope encompasses both "breadth," i.e., all subdimensions of the construct captured in the hierarchy are represented

in the constructs, and "depth", i.e., for any possible subdimension, all possible actionable causes of that subdimension are also explored and captured.

To test the framework, the tree will be transformed into a Computer-Adaptive Surveys (CAS), which are multi-dimensional instruments where large constructs with several interdependent dimensions are "unpacked" and mapped into a tree. In a CAS tree, each top-level construct represents a theory and should capture the entire scope of the problem domain, such as why Kanban is unsuccessful in organizations. In CAS respondents perform a depth-first traversal of the tree, where each stage of the traversal involves the respondent rating all items in the stage. Respondents then receive only child constructs associated with the lowest or highest rated constructs. Each respondent could traverse the CAS tree in a different way. To demonstrate, if self-efficacy is the area the user is least confident with, CAS retrieves questions on constructs such as organizational support or prior experience. If the user is least happy with organizational support, CAS then retrieves questions about workshops and development programs. CAS does not retrieve further questions on constructs the users rated satisfactorily. As the respondent continues to answer questions, CAS navigates deeper down the tree and questions roll down until the respondent hits one set of constructs with no children, hence in this example the level of expertise of the workshops is the area users are not happy with. For instance, if 30-40% of our users are telling us that expertise is a problem then this also means that Self-Efficacy is the theory that is driving our understanding of what the problem is. Of course, not every respondent would navigate the tree the same way. Thus, aggregating the results from respondents allows the researcher to observe the multiple major problems across respondents.

As future research, I intend to develop a CAS instrument to identify the principal root cause for why Kanban is unsuccessful in organizations. This study will have several significant contributions. One, the framework itself is a theoretical contribution as according to Doty (1994), configuration theories such as typologies and taxonomies, are more than theoretical

classification systems, as they have constructs, established relationships between the constructs and make predictions (Gregor, 2006). Similarly, a CAS tree, unpacks items to cover all the possible reasons related to the framework being studied and establishes a relationship among the constructs. Two, the results of the CAS will identify the root cause of why Kanban is unsuccessful, which is also a contribution to practice. Three, in addition, as I perform the CAS, I am likely to discover (and remedy) other impediments to Kanban adoption in software companies. Those will likely be a report for the software company. Finally, from a practice perspective, the framework, validation through surveys and incidental findings will help companies all over the world to become more efficient. In addition, the framework will identify potential areas for software development, this will help software companies to improve their productivity and creativity.

5.2.2 Example 2

Self-efficacy is key in understanding how individuals adopt new tools and easily develop considerable skill in the use of those tools. It is a pivotal concept that helps our understanding of technology acceptance, implementation, and use (Compeau & Higgins, 1995; Torkzadeh & Van Dyke, 2002). Bandura (1986, p 391) defines self-efficacy as, "People's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances. It is concerned not with the skills one has but with judgments of what one can do with whatever skills one possesses." Thus, people who have low self-efficacy would be less likely to perform related behavior in the future, in this case, adopt and use the mobile app, than those with high degree of self-efficacy (Keith, Babb, Lowry, Furner, & Abdullat, 2015).

There can be many reasons for why one has low-self-efficacy when using a tool and understanding and identifying these areas can potentially increase effective use (Compeau & Higgins, 1995). Computer-Adaptive Surveys (CAS) can help identify which of these issues is the most pressing, as solving the principal problem often times can resolve other issues. In

CAS, large constructs are unpacked to explore and identify the different dimensions. Each of these dimensions is represented in an item, where each item in CAS reflects a potential root cause. These items are then mapped in a CAS tree. Items concerning higher level concepts link to items with greater detail. The CAS tree is then transformed into a survey to test which of these items are most relevant to the context of the study.

For instance, consider an example on Instagram which is an online, mobile phone photo-sharing, video sharing, and social network service (SNS) that enables its users to take pictures and videos, and then share them on other platforms. The CAS tree is built in the following way. First, a CAS on Instagram self-efficacy would have a CAS tree with each top-level construct representing a different dimension of self-efficacy. The notion of scope encompasses both "breadth," i.e., all subdimensions of the construct captured in the hierarchy are represented in the constructs, and "depth", i.e., for any possible subdimension, all possible actionable causes of that subdimension are also explored and captured. Thus, there are four top-level constructs, perceived Instagram skill, confidence in ability to successfully find information online, level of Instagram content production, and level of Instagram content consumption. Each of these top-level constructs are unpacked to represent a dimension. As an example, perceived Instagram skill is unpacked to using and managing Instagram. Using Instagram is then unpacked to other social networks, use and functions, media and editing account management and privacy, and interactions and contacts. Each of these top-level constructs, then unpacks to numerous constructs.

Second, two blind and independent raters map the constructs into a hierarchy. Three, 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 nodes in the hierarchy. The hierarchies are then successively transformed to test if they branch in the same way. I found that the 3 measures of association, Goodman and Kruskal's Lambda, Cohen's

Kappa, and Goodman and Kruskal's Gamma together provide a "good enough" threshold(s) for assessing the overall similarity between tree hierarchies and diagnosing causes of disagreements between the tree hierarchies. Finally, to test the framework, the hierarchies are transformed into a survey.

This study will have several significant contributions. One, the framework itself is a theoretical contribution as a CAS tree, unpacks items to cover all the possible reasons related to the framework being studied and establishes a relationship among the constructs.

Two, the results of the CAS will identify the most pressing issues when using Instagram. Finally, the results of this study can improve the overall performance of Instagram.

5.3 Future work

As future research, I hope to explore and evaluate CAS in several areas. One is to assess CAS in other fields. As an example, studies have shown that data analytic tools such as Kanban have many benefits such as managing flow in teams and completing tasks more efficiently(Ahmad, Markkula, & Oivo, 2013). However, some end-users have been unwilling to use such tools. There could be many reasons for resistance, such as lack of understanding of key concepts of the tool (Ahmad et al., 2013) or end-users do not believe those tools will increase their performance. While there has been substantial research on usefulness and ease of use, little research has attempted to unpack the characteristics of technologies that make them easy to use or useful (Bagozzi, 2007; Legris, Ingham, & Collerette, 2003). I intend to use CAS to determine the root cause of why end-users are unwilling to use a particular data analytic tool. The outcome can assist both the organizations and software developers, such as increasing flow (Petersen & Wohlin, 2011).

Two, in this study, I was not able to fully assess the conclusion validity of CAS, as I found online reviews did not go into the depth that CAS trees go in. Hence, an additional method for assessing CAS results is required.

Three, this work is particularly useful for assessing the construct and content validity of two CAS trees. Future research is required for measuring other types validities for CAS, such as internal validity.

Fourthly, I intend to compare several popular measures used to compare trees with my statistical method to further demonstrate the use of this study's method.

Finally, this study used an adaptive version of branching for respondents to move from one set of items to another set according to a pre-defined criterion. I intend to assess and explore other possible options for determining the next set of question(s).

References

Anderson, D., Concas, G., Ilaria Lunesu, M., & Marchesi, M. (2011). Studying lean kanban approach using software process simulation. *12th International Conference*, XP 2011, 12–26.

Ahmad, M. O., Dennehy, D., Conboy, K., & Oivo, M. (2018). Kanban in software engineering: A systematic mapping study. *Journal of Systems and Software*, *137*, 96–113.

Ahmad, M. O., Markkula, J., & Oivo, M. (2013). Kanban in software development: A systematic literature review. In *In Software Engineering and Advanced Applications (SEAA), 2013 39th EUROMICRO Conference on* (pp. 9–16).

Ahmad, M. O., Markkula, J., Oivo, M., & Kuvaja, P. (2014). Usage of kanban in software companies. *In 9th International Conference on Software Engineering Advances.*

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, *8*(4), 244–254.

Balve, P., Krüger, V., & Tolstrup Sørensen, L. (2017). Applying the kanban method in problem-based project work: a case study in a manufacturing engineering bachelor's programme at Aalborg University Copenhagen. *European Journal of Engineering Education*, *42*(6), 1512–1530.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ, 1986*

Barreda, A., & Bilgihan, A. (2013). An analysis of user-generated content for hotel experiences. *Journal of Hospitality and Tourism Technology*, *4*(3), 263–280.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303–316.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1983). Beyond external validity. *Journal of Consumer Research*, *10*(1), 112–114.
Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, *19*(2), 189–211.
Dennehy, D., & Conboy, K. (2016). Early adoption of flow artefacts in ISD : An activity theory perspective. In *Thirty Seventh International Conference on Information Systems*.
Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*(1), 21–36.
Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218.
Diamantopoulos, A., & Winklhofer, M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277.
Doty, D. H. (1994). Typologies as a Unique Form of Theory Building : Toward Improved Understanding and Modeling Author (s): D . Harold Doty and William H . Glick Source : The Academy of Management Review , Vol . 19 , No . 2 (Apr ., 1994), pp . 230-251 Published by : Acad. *Academy of Management Review*, *19*(2), 230–251.
Freeze, R., & Raschke, R. (2007). An assessment of formative and reflective constructs in IS research. *ECIS 2007 Proceedings.*, 1481–1492.
Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., & Sadeh, N. (2013). Why people hate your app. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 13*, 1276.
Galesic, M., & 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.
Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611.
Jindal, N., & Liu, B. (2007). Analyzing and Detecting Review Spam. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 547–552
Keith, M. J., Babb, J. S., Lowry, P. B., Furner, C. P., & Abdullat, A. (2015). The role of mobile-computing self-efficacy in consumer information disclosure. *Information Systems Journal*, *25*(6), 637–667.
Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management*, *40*(3), 191–204.
Mackenzie, S. B., Podsakoff, P. M., & 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.
Markus, M. L. (2004). Technochange management: Using IT to drive organizational change. *Journal of Information Technology*, *19*(1), 4–20.
Maruping, L. M., Venkatesh, V., & Agarwal, R. (2009). A control theory perspective on agile methodology use and changing user requirements. *Information Systems Research*, *20*(3), 377–399.
Morse, Janice, M. (2000). Determining sample size. *Qualitative Health Research*, *10*(1), 3-5.
Petersen, K., & Wohlin, C. (2011). Measuring the flow in lean software development. *Software: Practice and Experience*, *41*(9), 975–996.
Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, 17(1), 2-26.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental for Generalized Designs Causal Inference*. *Wadsworth Cengage learning*. Houghton Mifflin Company.

Somprasertsri, G., & Lalitrojwong, P. (2010). Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. *Journal of Universal Computer Science*, *16*(6), 938–955.

Tanner, M., & Dauane, M. (2017). The use of kanban to alleviate collaboration and communication challenges of global software development. *Issues in Informing Science and Information Technology Education*, *14*, 177–197.

Torkzadeh, G., & Van Dyke, T. P. (2002). Effects of training on Internet self-efficacy and computer user attitudes. Computers in Human Behavior, 18(5), 479–494.

Wiener, M., Mähring, M., Remus, U., & Saunders, C. (2016). Control configuration and control enactment in information systems projects: Review and expanded theoretical framework. *MIS Quarterly*, *40*(3), 741–774.

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems*, *50*(4), 743–754.