EVALUATION OF THE PSYCHOMETRIC QUALITY AND VALIDITY OF A STUDENT SURVEY OF INSTRUCTION IN BANGKOK UNIVERSITY, THAILAND

by

Waritsa Chamoy

B.Ed., Secondary Education, Chulalongkorn University, 2006

M.Ed., Educational Measurement and Evaluation, Chulalongkorn University, 2008

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Waritsa Chamoy

It was defended on

July 24, 2018

and approved by

Clement A. Stone, PhD, Professor, Psychology in Education

Lauren Terhorst, PhD, Associate Professor, School of Nursing

Debra W. Moore, PhD, Instructor, Instruction and Learning

Dissertation Advisor: Suzanne Lane, PhD, Professor, Psychology in Education

EVALUATION OF THE PSYCHOMETRIC QUALITY AND VALIDITY OF A STUDENT SURVEY OF INSTRUCTION IN BANGKOK UNIVERSITY, THAILAND

Waritsa Chamoy, PhD

University of Pittsburgh, 2018

The main purpose of this study was to conduct a validation analysis of student surveys of teaching effectiveness implemented at Bangkok University, Thailand. This study included three phases; survey development, a pilot study, and a full implementation study. Four sources of validity evidence were collected to support intended interpretations and uses of survey scores. To this end, this study evaluated the extent to which the content evidence supported the construct definition of the survey (RQ1), the relationships among survey items and survey components corresponded to the construct dimension (RQ2), the survey exhibited gender differential item functioning (RQ3), and student ratings and a similar measure of teaching quality and student achievement (RQ4) were related.

Overall, the student survey demonstrated good psychometric quality and the intended purposes and uses of the survey were supported. Based on expert reviews, the dimensions and survey items were perceived adequate in covering teaching quality, the survey items were perceived to properly assess the associated dimensions, and the response scales were perceived suitable with what was intended to measure. Exploratory factor analysis suggested that the construct of teaching effectiveness as defined in this survey may be unidimensional. Although the results did not support multidimensionality, the dimensions can still be used by individual instructors to evaluate their own teaching. Cronbach's α coefficients were high and supported the internal consistency of the survey. There was no occurrence of gender DIF in this student survey. Therefore, the validity evidence of survey score interpretations was supported since the meaning of survey categories/scales was shared across male and female students. Finally, the results based on relation to other variables showed a strong positive relationship between the student survey and another currently used survey at Bangkok University which was used to evaluate teaching effectiveness for a decade. This could indicate that the student survey was measuring a similar construct of teaching effectiveness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This work would not have been done without all valuable guidance, help, and support from those of you who deserve to be acknowledged here. First of all, I would like to express my sincere gratitude to my advisor, Dr. Suzanne Lane, who had been my driving force towards the success of this dissertation. Every detail of your mentorship and every single word of your encouragement were extremely helpful and engraved in my mind. Your expertise in measurement has inspired me to conduct high quality work. Thank you for being patient with me throughout my long expedition in the Research Methodology program.

I would like to extend my greatest acknowledgement to my committee members, Dr. Clement A. Stone, Dr. Lauren Terhorst, and Dr. Debra W. Moore, for your professional guidance, strong expertise, and useful feedback. Thank you for sharing your experience and time with me. Your dedication and support had been invaluable to my success. I would not have been able to complete this work without all of you.

In addition, I gratefully acknowledge the financial support received towards my doctoral study from Bangkok University, Thailand. Special thanks to Dr. Sountaree Rattapasakorn, Assoc. Prof. Dr. Natthaphob Nimpitiwan, and Asst. Prof. Raweewan Kaewwit for supporting my career goals and facilitating the student surveys in Thailand. The director and colleagues at Office of Educational Quality Standards, it has been my pleasure to work with you. Thank you for your support and assistance. I would also like to thank all of the academic leaders, instructors, psychometricians, and students who participated in this study. Your devotion were greatly appreciated.

To my cohort-mates at the University of Pittsburgh and friends in Thailand, thank you for your remarkable moral support, encouragement, and well-wishes. You were a part of my doctoral journey. Parichat, my 5-year roommate, deserves special appreciation. Thank you for your interesting thoughts and being my best friend.

Most importantly, I would like to thank my family: mom, dad, grandma, aunts, uncles, and cousins for your unwavering love and care across thousand miles. You were always there with endless support and encouraging words. I look forward to returning home and making all of you proud. My mom, Nanthicha, deserves the most of my gratitude. Thank you for listening to me in every step of the way. Your unconditional love, care, and encouragement always gave me strength to move on in times of struggle. You, indeed, are behind everything I have ever succeeded.

1.0 INTRODUCTION

Why measure effective teaching? Real improvement in teaching requires quality measurement. To improve teaching, make critical personnel decisions, and demonstrate the performance of an institution, high quality feedback based on valid and reliable assessments need to be given. But measuring teaching is difficult and no single measurement tool is likely to capture teaching effectiveness (T. J. Kane, Kerr, & Pianta, 2014). Therefore, teaching evaluation systems are developed in an effort to systematically evaluate teaching. In higher education, colleges and universities rely on students to provide responses about the quality of the teaching that they experience using student surveys (Ferguson, 2012).

1.1 STUDENT RATINGS OF INSTRUCTION

One widespread measure of teaching effectiveness is student ratings. In higher education, almost every institution all over the world uses student ratings as an evaluation component of the teaching system (Zabaleta, 2007). Most educators believe that students are important stakeholders in the process of gathering insight into the quality of teaching in a course. Student ratings are also considered as a viable and cost-effective measure. Online student-rating systems allow colleges/universities to quickly collect ratings from thousands of students and reduce costs.

Additionally, the administration of the surveys is easy and can be standardized (Meyer, Doromal, Wei, & Zhu, 2016; Sorenson & Johnson, 2003).

The question is, to what extent can student ratings be used for making fair and valid comparative judgements about the effectiveness of teachers, courses, departments, and institutions (Kwan, 1999)? There are numerous research studies examining what scores from student ratings of instruction represent. Some research studies have investigated the relationship between student ratings and the characteristics they were assumed to represent: good teaching, and by extension, good teachers. Some researchers claim that student ratings are the most direct indicators of teaching quality and effectiveness. Underlying all these challenges is the fundamental question of validity.

1.1.1 Quality assurance (QA) in higher education

Quality in higher education is not a simple one-dimensional notion about academic quality. It has been viewed as a multi-dimensional concept by stakeholders (Ong Chee Bin, 2016). The World Declaration on Higher Education for the Twenty First Century: Vision and Action (October 1998), Article 11, Qualitative Evaluation considers quality in higher education as "a multi-dimensional concept, which should embrace all its functions, and activities; teaching and academic programs, research and scholarship, staffing, students, buildings, facilities, equipment, services to the community and the academic environment. Internal self-evaluation and external review, conducted openly by independent specialists, if possible with international expertise, are vital for enhancing quality." (p.25) To develop, implement, maintain and improve the level of quality in higher education, a university needs to install a quality assurance system. The Regional Report of Asia and the Pacific (UNESCO, 2003b) defines quality assurance in higher education as “systematic management and assessment procedures to monitor performance of higher education institutions”. (p.9)

1.1.2 AUN-QA model for quality assurance

The establishment of ASEAN University Network (AUN) and AUN - Quality Assurance Network (AUN-QA) in 1995 and 1998, respectively, has initially led to a new framework of quality assurance in higher education. To assist universities in developing, implementing, maintaining, and improving the level of quality, AUN-QA common policies, criteria, and strategic plan were first developed in 2001. The AUN-QA guidelines and manual for the implementation of the guidelines were later endorsed in 2004 and 2006, respectively. Effective in 2011, the universities in the ASEAN region were encouraged to implement the AUN-QA guidelines and apply the AUN-QA standards and criteria in order to develop adequate internal quality assurance (IQA) systems at institutional and program levels. IQA ensures that an institute or program has policies and mechanisms in place to make sure that it meets its own objectives and standards (ASEAN University Network, 2011). According to the latest version of the guide to AUN-QA assessment (Ong Chee Bin, 2016), the AUN-QA model for an IQA system consists of the following areas: internal quality assurance framework, monitoring instruments, evaluation instruments, special QA-processes to safeguard specific activities, specific QA-instruments, and follow-up activities for making improvements. Course and teacher evaluation are parts of the evaluation instruments and each institute is encouraged to develop its own instrument that fits its institutional needs.

With regards to IQA at the program level, it focuses on quality of educational activities in terms of input, process, and output. It starts with stakeholders’ needs and these needs are formulated into the expected learning outcomes which drive the program. There are 11 criteria including expected learning outcomes, program specification, program structure and content, teaching and learning approach, student assessment, academic staff quality, support staff quality, facilities and infrastructure, quality enhancement, and output.

Recently, Thai universities have begun adopting AUN-QA criteria to develop their own instruments to promote their academic standards and the quality of teaching. While there are many ways to evaluate teaching quality, the focus of this study is the development and validation of a newly developed student survey of teaching effectiveness. From the literature, student surveys are widely used to assess teaching effectiveness for two main purposes: formative and summative evaluation. First, student surveys can be used to provide feedback to teachers about the effectiveness of their teaching. Second, student surveys can be used in conjunction with other information for faculty promotion and tenure decisions. Third, they can be used to demonstrate an institution's performance as a part of internal quality assurance processes (Galbraith, Merrill, & Kline, 2012; Richardson, 2005).

Overall, student surveys of teaching effectiveness under the context of AUN-QA need to be implemented. The way student surveys are developed and validated needs to meet professional standards. Validity evidence is needed to determine whether ratings generated by a survey truly represent the defined construct and in what degree evidence and theory support the interpretations of the ratings (Onwuegbuzie, Daniel, & Collins, 2009).

1.2 THEORETICAL FRAMEWORK

To ensure a unified concept of validity for student surveys of teaching effectiveness, definitions and concepts of validity are briefly summarized below, followed by a critical analysis of student surveys research using the validation framework.

During the late 1970s, the development of validity theory focused on a clear identification of different kinds of validity evidence to evaluate particular interpretations and uses of test scores (Kane, 2006). During the 1980s, there was a major revision of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) to reflect some changes about the concept of validity. Messick's benchmark chapter on validity in the third edition of *Educational Measurement* (1989) provided a comprehensive framework of validity. Messick (1989) adopted a unified model of validity and defined validity as "an evaluation of proposed interpretations and uses of test scores regarding their adequacy and appropriateness of inferences and actions". Subsequently, the definition of validity in the 2014 *Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014)* refers to "the degree to which evidence and theory support the score interpretations entailed by intended uses of tests". Critical to this definition is that, validity is not a property of a test or assessment. Instead validity is a characteristic of the meaning and interpretation of the assessment scores and any actions based on the assessment scores.

With respect to the validation of student surveys of teaching effectiveness, it concerns whether ratings generated by the student surveys truly represent the defined construct of teaching effectiveness and in what degree evidence and theory support the interpretations of the ratings

(Onwuegbuzie et al., 2009). In other words, it refers to the extent to which the responses from the student surveys correspond to other indicators of teacher and teaching quality. Thus, validity evidence of teaching effectiveness must be provided. This study is a validation study of a new survey of teaching effectiveness based on an argument-based approach, which was developed to facilitate the validation process as a unified conception taking the interpretations, uses, argument, claim, and validity evidence into account (M. T. Kane, 2006)

There are potentially five different sources of validity evidence to collect to support the interpretations and uses of student surveys as indicators of teaching quality. This study focuses on obtaining four sources of validity evidence (survey content, response processes and expert review of items, internal structure, and relations to other variables). Given that the remaining validity evidence based on survey consequences is equally important, future studies should address it.

The psychometric assessment of student surveys of teaching effectiveness in this study was conducted through classical test theory and can contribute to validity evidence. In addition to classical test theory, an exploratory factor analysis (EFA), a confirmatory factor analysis (CFA) and an exploratory structural equation modeling (ESEM) were proposed to explore the factor structure of the survey. ESEM is an integration of EFA, CFA, and SEM (Asparouhov & Muthén, 2009; Marsh et al., 2010; Marsh et al., 2009). Finally, the most commonly used non-IRT method for detecting DIF in ordinal items, the generalized Mantel-Haenszel (GMH) procedure was employed to examine if the survey items exhibited differential item functioning between male and female students.

1.3 STATEMENT OF THE PROBLEM

Due to the importance of student ratings as part of classroom improvement, personnel decisions, or quality assurance, one must provide validity evidence of teaching effectiveness for these purposes. Many constructed instruments used to evaluate teaching effectiveness have not been subjected to an investigation of their psychometric quality or the validity of the survey score interpretations and uses.

Bangkok University, one of the biggest private universities in Thailand has adopted student surveys as a measure of teaching quality for several decades. With almost twenty thousand students enrolled in Bangkok University each year, the data from student ratings can be comprehensive and useful for evaluating teaching effectiveness. Nonetheless, the development process of student surveys was outdated without consideration of reliability and validity. Also, there were some changes based on the new framework of quality assurance, AUN-QA. Therefore, the development and validation of a survey that met the requirements of the AUN-QA framework was needed.

A series of validation analyses were conducted on the new survey. The major assumptions and inferences in interpreting the results of student ratings of instruction and the associated sources of validity evidence needed to be clearly identified before the survey was used. This study focused on the survey development and validation as well as its psychometric properties.

The survey validation followed a framework that seeks to establish validity through multiple sources of validity evidence. Collecting validity evidence provided justification for the set of inferences that were intended to be drawn from scores yielded by a student survey. Four sources of validity evidence were accumulated, including validity evidence based on survey content, response processes and expert review of items, internal structure, and relations to other variables.

An exploratory factor analysis (EFA), a confirmatory factor analysis (CFA) and an exploratory structural equation modeling (ESEM) were proposed to explore the factor structure of the student surveys of teaching effectiveness. The generalized Mantel-Haenszel (GMH) procedure was used to examine the extent to which differential item functioning (DIF) occurs.

In sum, the purposes of this study were to conduct a validation analysis of student surveys of teaching effectiveness implemented at Bangkok University, to explore the dimensionality of aggregated student ratings, and to analyze DIF. There were four research questions that guided this study:

Research question 1: To what extent does the content evidence support the construct definition?

Research question 2: To what extent do the relationships among survey items and survey components correspond to the construct dimension?

Research question 3: Is there gender differential item functioning in student ratings?

Research question 4: Are there relationships between student ratings and a similar measure of teaching quality and student achievement?

1.4 SIGNIFICANCE OF THE STUDY

The current investigation aimed to develop and validate an instrument to measure teaching effectiveness using student ratings. This study has both theoretical and practical significance. In a theoretical sense, this study contributes to the existing literature related to the validation of student ratings of instruction and a better understanding of the characteristics of student ratings. In the practical sense, the study provides solid validity evidence for the applicability of student surveys of teaching effectiveness in Thai higher education. The intention of this study was to provide the university administrative staff with important information about the student surveys of teaching effectiveness as it relates to the validity of interpretation and use of results in improving classroom practices, making personnel decisions such as tenure and promotion, and demonstrating the institution performance for quality assurance purpose.

1.5 ORGANIZATION OF THE STUDY

This document is organized into 5 chapters. Chapter one provides some background information about student ratings of instruction and the validation framework. The purpose of the study, the research questions, as well as the significance of the study are also introduced in Chapter one. The next section, Chapter two, reviews the literature on student ratings as a measure of teaching quality as well as the issues in using ordinal scales in survey research. Following this is the conceptual foundations of validity. Any conceptual and methodological challenges raised by student surveys are also addressed in Chapter two. This chapter, also provides an overview of EFA, CFA and ESEM approaches proposed to explore the factor structure of the survey and DIF analyses using the generalized MH procedure. Next, Chapter three presents the overall design of the study including steps in survey development and validation. A pilot study and its results and a full implementation study are also detailed in Chapter three. Chapter four provides results from the full implementation study to answer the four research questions. The results from the pilot study are also summarized here. Finally, chapter five provides a summary and discussion of the study findings for each research question. Following this is the limitations of this study and important directions for future research.

2.0 LITERATURE REVIEW

The main purpose of this study was to develop and evaluate the psychometric quality and validity evidence of student ratings of instruction for a new survey. This survey was developed based on a thorough review of the literature. Four sources of validity evidence were obtained to support score interpretations and uses including evidence based on survey content, response processes and expert review of items, internal structure, and relations to other variables. Using Classical Test Theory (CTT), the study first examined the psychometric properties of the survey using a series of data analyses. Afterwards, EFA, CFA and ESEM were used to examine internal structure of the survey. Finally, the GMH method was employed to examine whether any items exhibit DIF between male and female students.

To cover the essential background for this study, the first section of the chapter describes the necessary information on teaching quality under the AUN-QA framework in the ASEAN region. The second section provides a brief history of student ratings. The issues in using an ordinal scale are considered next. Following a discussion on validity in student ratings of instruction, with an emphasis on four sources of validity evidence is reviewed. The next section is an overview of EFA, CFA and ESEM approach for examining internal structure of the survey, while the last section provides a description of detecting DIF using the GMH method.

2.1 TEACHING QUALITY UNDER AUN-QA FRAMEWORK

The AUN - Quality Assurance Network (AUN-QA) indicates that student surveys of teaching effectiveness need to be based on three criteria of AUN-QA including criterion 4: teaching and learning approach, criterion 6: academic staff quality, and criterion 11: stakeholders' satisfaction (students). All three AUN-QA criteria are used to facilitate the assessment of teaching quality and discussed below (Ong Chee Bin, 2016).

1. Criterion 4: Teaching and learning approach. This AUN-QA criterion focuses on how students acquire the following skills after being exposed to each teaching and learning approach:

The ability to discover knowledge for oneself - learners have research skills and the ability to analyze and synthesize the material they gather. Learners understand various learning strategies and can choose the most appropriate for the task at hand.

The ability to retain knowledge long term - an approach to learning that emphasizes construction of meaning rather than memorizing facts for greater retention.

The ability to perceive relations between old knowledge and new - quality learning is always trying to bring information from various resources together.

The ability to create new knowledge - quality learners discover what others have learned and documented, perceiving the relations between that knowledge and their own experiences and previous learning to develop new insights.

The ability to apply one's knowledge to solve problems

The ability to communicate one's knowledge to others - quality learners form and substantiate independent thought and action in a coherent and articulated fashion.

An eagerness to know more - quality learners are lifelong learners.

2. Criterion 6: Academic staff quality. This AUN-QA criterion focuses on the evaluation of academic staff competence. The quality of academic staff encompasses qualification, subject matter expertise, experience, teaching skills, and professional ethics. In other words, it is important that those who teach have a full knowledge of and understand the subject they are teaching. In addition, they need to have the necessary skills and experience to communicate their knowledge and understand students effectively in a variety of teaching contexts.

3. Criterion 11: Output (students' satisfaction). This AUN-QA criterion focuses on whether a university has a structured method for obtaining feedback from students in term of their satisfaction toward the quality of teaching. Specifically, it is concerned with what students think about the courses, teaching, examinations, assignments, and classroom activities.

2.2 HISTORY OF STUDENT RATINGS

Student ratings of instruction in higher education have been formally in existence since the 1920s. In 1920–1925, students at the University of Washington rated teaching quality on what was credited as being the first student rating form. In addition, the first empirical research on student ratings conducted by the researchers from Purdue University was published. Class size and lecture versus discussion were appeared on the research study. Some of the findings indicated that small class sizes were more effective than large class sizes in term of learning retention and the discussion teaching method was better than the lecture teaching method for long-term retention (McKeachie, 1990).

In the late 1960s and early 1970s, the use of student ratings expanded across American universities in response to students' calls for accountability. In the 1980s – 1990s, student ratings of instruction were used for faculty improvement and administrative purposes (Onwuegbuzie et al., 2009). Overall, student ratings of instruction can be used for both formative and summative purposes.

Some researchers were skeptical about students' ability to rate the quality of their teachers and courses and questioned the validity of such ratings. A review of empirical studies by Costin, Greenough, and Menges (1971) indicated that student ratings can provide reliable and valid information on the quality of teachers and courses. In late 1997, a series of research studies by d'Apollonia and Abrami (1997), Greenwald (1997), Marsh and Roche (1997), and McKeachie (1997) reviewed the literature on student ratings of instruction and their uses. For example, Marsh and Roche (1997) evaluated the validity and the usefulness of student ratings. They concluded that under appropriate conditions, student ratings were relatively valid against a variety of indicators of effective teaching and unaffected by a variety of potential biases (e.g., grading leniency, class size, and workload). Student ratings were useful in improving teaching quality but insufficient. Generally, all four studies favored the use of student ratings and confirmed that issues regarding the validity of student ratings had been settled.

A series of articles by Cohen (1980, 1981, and 1982) was conducted to provide more information regarding the validity of student ratings using multisection validity studies. Various research studies on the relationship between student ratings and student achievement across different courses and instructors were synthesized. They demonstrated that overall student responses on rating scales of teaching effectiveness are positively associated with their academic achievement. This result strengthens the claim that student ratings are valid measures of teaching effectiveness. It is important to be aware that a well-constructed and score-validated instrument of student ratings can be a useful indicator of teaching quality, as measured by student achievement (d'Apollonia & Abrami, 1997; Penny, 2003).

The continued use of student ratings as a measure of teaching quality needs to be based on empirical research pertaining to the validity and reliability. Student ratings are unlikely to be used in administrative decisions such as promotion and tenure, but there is sufficient agreement that they are useful for teaching improvement purposes (Penny, 2003). Most studies on teaching improvement derives from studies of the short-term effects of feedback from student ratings of instruction (Chatterjee, Ghosh, & Bandyopadhyay, 2009). It is to be noted that teaching improvement and course improvement have the same goal, and the former refers to the activities the instructor designed to facilitate student learning, while the latter refers to broad choices in course content and to the general structure of the course.

Most of the student surveys use a rating scale such as a four, five, or six-point ordinal scale to measure students' perspectives and reactions. The next section discusses issues regarding the use of ordinal scale.

2.3 ISSUES IN USING ORDINAL SCALES

In the world of survey research in social science, data measurement theory becomes significant, where individuals respond to questions that are crafted for soliciting their qualitative observations (Granberg-Rademacker, 2010). An ordinal scale is widely used to obtain this kind of data. In ordinal-level measurement, the categories must be homogeneous, mutually exclusive, exhaustive, and ordered along some continuum. For example, the respondents are asked to select one category representing their frequency toward a particular statement running from, *never* to *always* (i.e., never, rarely, sometimes, very often, always), and it is possible to make such statements as "*sometimes*" is more frequent than "*rarely*". Nevertheless, it is not possible to say how frequent "*sometimes*" is as compared to "*rarely*". It is assumed that the distance between categories is unequal. Therefore, differences in magnitude cannot be considered (Blaikie, 2003).

The Likert scale is one type of ordinal scale in psychometrics and is used extensively in the social sciences and educational survey research. It was devised to measure attitude without deriving item scale locations and values (J. S. Roberts, Laughlin, & Wedell, 1999). The logical property of the Likert scale is a passive/ selective response format within the affective domain (Carifio & Perla, 2007; Jamieson, 2004). In other words, it is a set of negative or positive statements (items) offered for a real or hypothetical situation under study. Participants, in general, are asked to indicate their level of agreement on any given question using some common categories: "*strongly disagree*" to "*strongly agree*". All statements in combination represent the specific dimension of the attitude towards the given situation (Joshi, Kale, Chandel, & Pal, 2015).

The Likert scale is analytically treated and interpreted depending upon it being a symmetric or asymmetric scale. The construction of a symmetric scale is when the position of neutrality lies exactly in between two extremes of strongly disagree to strongly agree. Thus, a participant is independent in selecting any response in either directions in a balanced and symmetric way. In contrast, there are less choices on one side of neutrality as compared to the other side for an asymmetric Likert scale. It sometimes refers to forced choices with no value of neutrality (Joshi et al., 2015).

There are several measurement and statistical issues associated with using an ordinal scale, especially the Likert response format in survey research. The issues related to reliability, validity, robustness, analyses, and number of categories are discussed in the next sections.

2.3.1 Reliability issues

The issues of reliability of scores are a basic psychometric property for any scale. For a measurement of any latent trait, reliability refers to the degree to which the observed individual differences are represented by true individual differences (Deng, Marcoulides, & Yuan, 2015). Generally speaking, the literature has shown that the estimates of coefficient alpha computed from ordinal response data are downward biased when compared with one computed from an interval/ratio scale. Nonetheless, when the theoretical reliability increases, the magnitude of bias decreases (Zumbo, Gadermann, & Zeisser, 2007).

When outliers are taken into account, Liu and Zambo (2007) found that the estimates of Cronbach's coefficient alpha for interval/ratio item response were seriously inflated by the outliers. Following this, Liu, Wu, and Zambo (2010) studied the impact of outliers on estimates of reliability for ordinal item response data, but the examination was more complicated because of the nonnormal distribution of ordinal data. A Monte Carlo simulation was performed to examine the impact of outliers on bias and efficiency of Cronbach's coefficient alpha. Similar to the result for interval/ratio item response, the impact of outliers increased the inflation of bias and efficiency of Cronbach's coefficient alpha for ordinal scale item response data. Nevertheless, when the theoretical reliability increased, the inflation decreased.

In addition, the influence of the number of Likert scale points on reliability estimates was examined. The study by Zumbo, Gadermann, and Zeisser (2007) concluded that the bias on reliability estimates decreased when the number of scale points increased. Also, the number of response categories had less of an effect on the theoretical reliability. When those two factors were combined (i.e., outliers and number of scale points), the number of response categories only inflated the efficiency of Cronbach's coefficient alpha when the number of options was increased in the presence of outliers. They found that fewer response categories are superior in the presence of outliers. Without outliers, more response categories are superior.

2.3.2 Validity issues

Validity issues arise because of the item scale itself. Specifically, an ordinal scale only has fixed response choices possible for individuals to choose, hence the individuals are expected to truncate or round off their level of agreement to the extent that it fits a scale. In this case, errors are produced (Granberg-Rademacker, 2010; J. S. Roberts et al., 1999)

Another validity concern is how individuals handle and interpret the scale. The different understanding of respondents toward the same item is often a problem for ordinal scales. This problem arises when data is analyzed and interpreted, because the meaning of coded values is not shared across subgroups. This concern can be reflected in Differential Item Functioning (DIF). To illustrate the DIF issue, the assumptions of survey research are that *X* is the interval or ratio-level latent variable and all respondents interpret the item responses and the items themselves in the same way (Granberg-Rademacker, 2010).

$$f(X) \approx x. \tag{2.1}$$

The notation in equation (2.1) does not need to be equal for ordinal level scales because of their preferential rankings. The equation (2.1) holds true when the understanding of the scale points of individual follows a general order of precedence. Let each subscript denote the possible scale points (e.g., r_1 is strongly agree, r_2 is agree). If the following notation holds true for all individuals, equation (2.1) is valid:

$$r_1 \prec r_2 \prec \ldots \prec r_k, \tag{2.2}$$

where k is responses running from k = 1, 2, … , K. If equation (2.2) holds true, then individual’s understanding of the monotonically increasing responses, r_k, is isomorphic. But if it does not hold true, then it represents the uniqueness of response understanding and interpretation as noted in equation (2.3):

$$f_i(X_i) \approx x_i. \tag{2.3}$$

If equation (2.3) holds true, then it is quite possible that the interpretation of the scale points themselves differs based on the individual i respondent’s interpretations such that $r_{j,i} \neq r_{j,-i}$, thus raising the possibility of encountering DIF. It is hoped that all response understandings are the same for every respondent as noted in equation (2.4):

$$f_1(.) \approx f_2(.) \approx \cdots \approx f_i(.). \tag{2.4}$$

If equation (2.4) holds true, then equations (2.1) and (2.2) also hold true, and there is no DIF problem. But if equation (2.4) does not hold true, then the researcher should consider corrective interpretative methods, such as anchoring scales or vignettes, to correct for DIF.

2.3.3 Robustness issues

Robustness is concerned with the chance of coming to an erroneous conclusion when the wrong statistical technique is used. If the chance does not increase much, robustness holds. This issue occurs frequently when using Likert item responses and an inappropriate statistical technique is adopted. Based on the review of Jamieson (2004), researchers needed to be cautious when employing the appropriate parametric tests for interval/ratio variables to ordinal variables because of the possibility of violating the normality assumption. Non-parametric tests were suggested to use for ordinal variables as the appropriate statistics. However, Norman (2010) argued that those non-parametric tests were seldom employed because they can only handle the simplest designs. This issue was carefully reviewed by Norman (2010) and the following conclusions were drawn from his review of the literature. Parametric tests assume that the distribution is normal, but it is the distribution of means, not of the data. The study of their robustness found that they were robust even for non-normal distributions with high skewness. Particularly, for sample sizes greater than 5, the assumption of normality is not required. Hence, using parametric tests for ordinal variables is justified. For example, Norman (2010) explained that correlation inherently deals with variation, not central tendency and the correlation magnitude is sensitive to extreme distribution. However, using ordinal response scales with less than 5 response categories rarely produces extreme distribution because the range of categories is narrow. This conclusion also gets supported from the empirical study. Correlation coefficients can be calculated for any dataset and they are robust with respect to violations of assumptions (Norman, 2010).

2.3.4 Analysis of the ordinal scale issues

The issues of ordinal scale analysis are discussed by beginning with the question – which type of Likert scale? There are two schools of thoughts: one school considers Likert scales as ordinal and another treats it as interval. The first focuses on the ranking order of Likert scales as well as the unequal distance between each pair of two response categories. Hence, it is considered an ordinal scale. The other focuses on the researcher's aim to combine all the items in order to generate a composite score for a respondent because a single Likert item is rarely analyzed. It has been argued that it is defensible to indicate that a Likert scale, as a sum of all items, is interval (Joshi et al., 2015).

One problem with ordinal data in many statistical procedures is that the values cannot assume to be normally distributed. Thus, the bias may be produced when analyzing an ordinal scale. However, as discussed earlier for the robustness issue, some of the parametric tests are robust with respect to violation of the normality assumption. Hence, when all items are combined together, the results from the analysis of ordinal data using parametric tests suggest that they are robust but have less statistical power compared to their nonparametric competitors (Harwell & Gatti, 2001; Norman, 2010). In sum, the analysis of ordinal scale aims to identify the most discriminating, homogeneous, and reliable items regardless of any nonparametric or parametric test selected (J. S. Roberts et al., 1999).

2.3.5 Number of categories issues

Another consideration is the number of options used for the score scale. Several studies have concluded that the appropriate number of response categories ranges from four to seven categories. An odd number of categories allow researchers to include a neutral position in the scale points (i.e., "*neither agree nor disagree*"). On the other hand, if researchers intend to elicit attitudes or opinions of respondents through their responses of "agree" or "disagree" without including the choice of neutrality, an even number of categories is preferable (Joshi et al., 2015; Wakita, Ueshima, & Noguchi, 2012).

Recent studies examined the impact of the different numbers of points on the Likert scale. Joshi et al. (2015) suggested that a 7-point Likert scale performed better than 5-point Likert scale because it provided more options which in turn increased the probability of accurately capturing respondents' beliefs. The study of Preston and Colman (2000) showed that a 2-point to 4-point Likert scale provided the lowest test-retest reliability, while a 7-point Likert scale provided the greatest item-reliability. Also, a simulation was conducted and found that when the number of options was less than eight, the interrater reliability increased (Cicchetti, Shoinralter, & Tyrer, 1985).

Another issue regarding number of scale points is its impact on the psychological distance between categories. Psychological distance is the distance between categories of ordinal scales. It is assumed that if the psychological distance between categories is equal, the scale will provide exact measurement of the psychological trait being assessed (Wakita et al., 2012). Wakita, Ueshima, and Noguchi (2012) examined whether the different number of categories affect the psychological distance between categories. The 4-, 5-, and 7- point Likert scales were included in the study. The results indicated that the psychological distance between categories was affected by the different number of categories. First, an increase in the number of options biased respondents against answers containing the strongest expressions. Second, the psychological distance deviated more as the number of categories increased. Specifically, the 7-

point Likert scale deviated more than the 4- and 5- point scales. This result suggested that the 7-point Likert scale was not recommended to adopt due to the sensitivity of the psychological distance.

2.4 VALIDITY OF STUDENT RATINGS OF INSTRUCTION

Student ratings of instruction are used for three purposes: (a) providing feedback to teachers for instructional improvement, (b) providing input for administrative decision making, that is faculty promotion and tenure decisions, and (c) providing evidence for demonstrating the performance of an institution as a part of internal quality assurance processes (Galbraith et al., 2012; Kember, Leung, & Kwan, 2002; Richardson, 2005; Spooren, Brockx, & Mortelmans, 2013).

The question is, to what extent can student ratings be used for making fair and valid comparative judgments about the effectiveness of teachers, courses, departments, and institutions (Kwan, 1999)? To answer the question, two kinds of arguments need to be developed according to Kane (2006); an interpretation/use argument (IUA) and a validity argument. The IUA can be considered as a clear statement specifying the proposed interpretations and uses of scores from student surveys together with its inferences and assumptions. The inferences and assumptions are derived from the intended interpretations and purposed uses of scores. Whereas, the validity argument is defined as an evaluation of the proposed IUA. To claim that a proposed interpretation or use is valid is to claim that the IUA is clear, coherent, and complete. That is, its inferences are reasonable and its assumptions are plausible. For developing those arguments used to support the intended interpretations and uses of scores from student ratings of instruction, an explicit structure of where the argument may be based is required.

First, the assumptions are needed to be explicitly stated regarding the meaning of student ratings of instruction. Specifically, the focus is on the relationship between scores from student ratings of instruction and the potential consequences of using these scores to indicate “effective” or “ineffective” teaching. To frame the assumptions, both actual and proposed uses of student ratings of instruction should be considered (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The assumptions and related inferences for student ratings of instruction are outlined below based on Kane’s (2006) four scoring inferences (Bell et al., 2012; Hill, Kapitula, & Umland, 2011); adapting from Kane, 2006).

There are four major assumptions and inference types in interpreting the results of student ratings of instruction and the associated IUA: scoring, generalization, extrapolation, and implication. The IUA are framed to make certain that they go through the three main purposes of the use of student ratings of instruction as discussed earlier.

First, the scoring inference (from observed performance to observed score) employs a scoring criterion to assign each teacher’s performance score on their teaching effectiveness. This inference relies on scoring criteria such as a Likert scale with five different score scales, semantic differential scale with a 7-point scale of opposite-meaning terms, etc. Those scoring criteria connect the inference into observable performance as well as provide information for evaluating teacher performance. The assumptions associated with the scoring inference include the scoring rules are suitable with what is intended to measure in the classroom environment, they are appropriately applied to each classroom context, and free of bias. In addition, the data need to fit the model.

To support this inference, the evidence that demonstrates how students encounter and process their understanding to items and scales on the student survey is needed. Evidence from research studies should,

a. Show that the response scale is appropriate, and the students are expected to apply the full range of the response scale to the performances in the classroom
b. Show that students understand the survey item and its scale in the same way.

Second, the generalization inference (from observed score to universe score) considers the representative of student ratings of the domain of teaching. For example, if the scores are used for faculty promotion and tenure decisions, it is important that the scores can be generalized from one specific course to all courses that the faculty teach in a semester. For student ratings of instruction, we only want to account for factors that shape teaching quality. However, there are several sources of variation regardless of the generalization inference that can affect samples of teaching quality, including the length of the survey, the administration date and time, the data collection process, the number of lessons in each course. Thus, it is suggested that those features need to be examined to evaluate the degree to which they have been accounted for. There are two assumptions associated with the generalization inference, that the sample of student ratings is representative of overall teaching quality and large enough to control sampling error.

With respect to validity evidence, it is important to examine evidence about how representative the survey items are of the domain of teaching effectiveness. In other words, the rating items must cover and represent the processes, strategies, and knowledge domain of teaching quality. The construct domain of teaching quality needs to be defined and investigated. Particularly, the item content from student ratings should;

a. relate to the characteristics or behaviors of good teaching derived from literature or from expert judgments, and

b. not reflect construct underrepresentation or/and construct irrelevance.

Third, the extrapolation inference (from universe score to target score) connects scores from student ratings of instruction to the broader teaching quality domain. If the protocol of student ratings of instruction is based on an extensive construct of teaching quality, strong evidence is needed to support that scores from student ratings reflect teaching quality. Student ratings with other measures of teaching effectiveness should be compared to examine their relationships. The extrapolation inference also considers the systematic errors that occur as the distance between scores from student ratings of instruction and the broader concept of teaching quality. Two assumptions associated with the extrapolation inference are that the competencies developed in the courses are required in responding to the statements on the teacher survey and no irrelevant sources of variability affects the interpretation of student ratings. Scores from student ratings of instruction should derive from the influence of teacher characteristics and teaching quality on student ratings. Therefore, the validity evidence to support this inference is that the correlation between scores from student ratings and other indicators of teacher and teaching quality should be more related than the correlation between hypothetically unrelated constructs. Particularly, scores from student ratings should;

a. correlate with expert ratings of teaching quality

b. correlate with estimates of teachers' knowledge

c. fail to correlate with unrelated constructs, such as the students' characteristics in a classroom.

Finally, the implication inference (from target score to verbal description) involves the impact on classroom practice, faculty tenure and promotion, and institution's performance. For example, one might evaluate whether a teacher who makes changes to classroom practices by employing the findings from the student ratings actually improves the overall quality of teaching, or another purpose is to investigate whether the consequences of the faculty who receive low student ratings are appropriate. The assumptions associated with the implication inference are that the implications of the student ratings are appropriate and their scores are stable across various situations, that is different students, teaching styles, schools, semesters, and years.

The validity evidence to support this inference for the purpose of educational improvement is that the use of student ratings of instruction in accountability decisions must not create negative consequences for stakeholders. Decisions based on scores from student ratings must;

a. identify both excellent and poor teachers with a reasonable degree of accuracy, and
b. not distort incentives for educators working within the system.

As discussed earlier, it is obvious that each of these inferences associates with various assumptions along with descriptions of validity evidence that can support the inferences. Considering the validity evidence, the validity evidence based on response processes and expert review of items can be used to support the scoring inference and its assumptions. While, the validity evidence based on survey content can be used to the support generalization inference and its assumptions. The validity evidence based on relationships to other variables can be used to support the extrapolation inference and its assumptions, and lastly, the validity evidence based on consequences of testing can be used to support the implication inference and its assumptions.

All validity evidence discussed above are based on the 2014 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Five sources of validity evidence and their issues and concerns in supporting a validity argument of student ratings of instruction are discussed in detail later. It is important to note that this study only focused on obtaining four sources of validity evidence (survey content, response processes and expert review, internal structure, and relations to other variables). This section focuses on the explicit statements of the claims being made that are used to evaluate those intended interpretations and uses in the IUA. Each of the four arguments above is also related to four claims and these four claims are used in specifying the IUA and validity arguments. Four claims: scoring, generalization, extrapolation, and implication claims are discussed in detail below (Bell et al., 2012; adapting from Kane, 2006).

1) The scoring inference makes claims about whether the scores from student ratings of instruction are suitable, accurate, reliable, bias free, and support the scoring model fit.

2) The generalization inference makes claims about whether the sample of student ratings of instruction is representative of the universe of generalization, that is the domains of teaching quality.

3) The extrapolation inference makes claims about whether the scores from student ratings of instruction are related to the target domain, which can be other measures drawn from the teaching quality domain. Additionally, it is important to claim that no systematic errors affected the extrapolation.

4) The implication inference makes claims about whether the scores from student ratings of instruction actually improve the quality of teaching and appropriate for two main purposes.

Specifically, it is necessary to claim that a measure of student ratings of instruction is a useful tool for making changes to classroom practices and making administrative decisions about faculty.

Together, the assumptions and claims are tested using empirical evidence. Sound and relevant evidence are developed for assessing the inferences and supporting assumptions entitled by the proposed interpretations and uses of student ratings (M. T. Kane, 2006). The five different sources of validity evidence are discussed below as well as their surrounding issues and concerned are also examined. Note that the different aspects of validity are illuminated by these evidence sources. Nevertheless, owing to a unitary concept of validity, these five sources of validity evidence do not indicate different kinds of validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

2.4.1 Collecting validity evidence based on survey content

Is there a relationship between the content of the student surveys and the target domain intended to be measured? Evidence can be obtained from the logical or empirical analyses of the adequacy of the survey content in representing the target domain and of the relevance between the target domain and the intended score interpretations and uses (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). This source of validity evidence can support the generalization inference. To make valid inferences about student ratings of instruction, the experts' judgments which are the logical analysis of the examination of the relationship between the content of student ratings of instruction and the construct being measured must be provided. The similarities between item content and the characteristics or behaviors of good teaching derived from the literature must be established. Additionally, evidence is obtained from the experts' judgements regarding an appropriateness and representativeness of the given set of items in the instrument. Thus, a clear definition of effective teaching in terms of the measurable characteristics or behaviors is important (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Ory & Ryan, 2001). However, in gathering this evidence, there are some issues/concerns needed to consider and discuss.

First, several validity studies of student evaluation of teaching concluded that many institutions lack a clear theory to support the operational definition of effective teaching. Consequently, the item content is possibly flawed and the validity evidence based on survey content is threatened (Onwuegbuzie et al., 2009; Ory & Ryan, 2001; Penny, 2003). Another issue concerns the differences in the meaning of effective instruction across subgroups of respondents. In other words, the respondents tend to respond to survey questions based on their own conceptions (Goldstein & Benassi, 2006; Kember & Doris, 2011; Kember & Wong, 2000; Onwuegbuzie et al., 2007). Specifically, the study of Onwuegbuzie et al. (2007) found that the students' perceptions toward important characteristics of good instruction do not match with what the survey developers deem to be their target traits. Moreover, students with different demographic backgrounds such as gender, year of study, and major treated the survey items differently. This is consistent with the findings of Goldstein and Benassi (2006) and Kember and Doris (2011) in that students' perceptions of good and poor teaching are commonly varied and influenced by students' beliefs which result in systematic variation. Additionally, the study of Kember and Wong (2000), pointed out that effective teaching is conceived by students in four categories (i.e. knowledge, learning, understanding, and beliefs) which are the quadrants formed by the intersections of the representations of beliefs about learning and perceptions of teaching. This distinction eventually suggests that students' criteria in responding to the survey of teaching effectiveness can be biased by the students' conceptions of learning. The final issue concerns the representativeness of the instrument items with respect to effective teaching. Some survey items are not measuring the constructs that they are supposed to measure (Marks, 2000; McKone, 1999; Onwuegbuzie et al., 2009). For example, the study of Marks (2000) found that student evaluations only represented students' perceptions and impressions which raised a specific concern of construct-underrepresentation.

2.4.2 Collecting validity evidence based on response processes and expert review

Does the nature of the student rating process fit the construct being measured? This evidence can provide the fit between the construct being measured and the detailed nature of response actually engaged in by survey takers. It is what students do when they encounter the survey item and process their understanding to the item. This source of validity evidence can support the scoring inference. To make valid inferences about student ratings of instruction, observers' agreement, a detailed description of assessment, expert review of items, and evidence from research studies need to be obtained (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Kane, 2006). For the student evaluation instruments, the students are asked to respond in terms of how strongly they agree or disagree with a set of statements. How those students react to questions is a main point to consider with regard to response processes.

2.4.3 Collecting validity evidence based on internal structure

To what extent do the relationships among survey items and survey components correspond to the construct domain? The correlations among domain and dimension scales can be analyzed to support this kind of evidence (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Kane, 2006). The conceptual framework for teaching effectiveness can be considered either unidimensional or multidimensional. For instance, a global score from a unidimensional survey is needed as a measure of overall teaching effectiveness when the student ratings are used for personnel decision making. On the other hand, the scores from several dimensions are needed to provide feedback to faculty in terms of specific areas in need given that teaching quality consists of several aspects. However, the multidimensionality of teaching is widely accepted and weighted averages of specific dimensions can be generated as an overall measure of teaching competency (Spooren et al., 2013).

Most of the recent studies of internal structure focused on the relationships among several items and dimensions by presenting item-total correlations and item-dimension (subscale) correlations. A couple of studies were concerned with the definition and operationalization of some teaching constructs. In this case, some survey items were not written in a way that was consistent with their corresponding domain. In the analysis, they found that the correlations of the survey items that belong to the same underlying domain yielded less of a relationship than those belonging to other dimensions (Bell et al., 2012; Spooren et al., 2013). This, at least, presents some evidence relevant to validity.

2.4.4 Collecting validity evidence based on relations to other variables

Is there a relationship between student ratings and variables external to the rating forms that are (1) expected to be predicted by the ratings and (2) similar measures of the intended construct? This source of validity evidence can support extrapolation inference and is comprised of convergent and discriminant evidence, test-criterion relationships, and validity generalization (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

First, convergent evidence is provided by the relationships between student ratings and other measures of instructional quality. The most common method to collect this evidence for student ratings of instruction is to obtain the relationships between student ratings and other measures of student achievement or learning (i.e., grades, test performance) as well as measures of effective instruction from alumni ratings, peer ratings, or self-ratings (Braun & Leidner, 2009; Clayson, 2009; Marsh & Roche, 2000; Onwuegbuzie et al., 2009; Richardson, 2005; Spooren et al., 2013). These relationships are helpful in enhancing the meaning of score interpretations and uses and positive relationships are expected. Onwuegbuzie et al. (2009) examined the relationship between student ratings and achievement and a positive effect was present in the finding. This effect implies that students who achieve high scores on achievement test tended to rate their teachers high.

In contrast, discriminant evidence is provided by the relationships between student ratings of instruction and measures intended to assess different constructs other than instructional quality. Discriminant evidence to support intended interpretations and uses of student ratings of instruction can be obtained by the relationships to other variables that reflect possible sources of bias or unwanted influence. This evidence involves the relationships between student ratings and some characteristics that are not directly related to teaching effectiveness including the following three characteristics;

1) Course characteristics (e.g., class size, level, teaching approach, workload) (Beran & Violato, 2005; Koon & Murray, 1995; Kwan, 1999; Marsh & Roche, 2000)

2) Instructor characteristics (e.g., gender, academic degree, teaching experience) (Centra & Gaubatz, 2000; Kohn & Hatfield, 2006; Wigington, Tollefson, & Rodriguez, 1989)

3) Student characteristics (e.g., gender, age, year in school, prior interest, expected grade) (Eiszler, 2002; Kohn & Hatfield, 2006; Marsh & Roche, 2000; Wachtel, 1998; Worthington, 2002)

A number of studies have been conducted to study the relationships of those various factors and student ratings of instruction as mentioned above, some of them revealed little interactions or no uniform patterns, while some of them showed significant bias in student ratings. For instance, recent studies of Centra and Gaubatz (2000), Kohn and Hatfield (2006) examined the impact of gender on student ratings of instruction. The results revealed that female students generally give their teachers higher ratings than male students. When considering instructor gender, the results also reflect some gender preferences, that is female students are more likely to rate male instructors high rather than male students. The study by Beran and Violato (2005) indicated that there is little variance in student ratings of instruction with respect to student and course characteristics (i.e., year and program of study, type of course, workload, expected grade). Similarly, another study found that students' expected grades are related to student evaluations of teaching to a small degree (Marsh & Roche, 2000).

Another form of evidence about relations to other variables is test-criterion relationships. This evidence of validity can be obtained from the correlation between student ratings and relevant criterion (Eiszler, 2002; Marsh & Roche, 1997; Spooren et al., 2013; Worthington, 2002). The choice of the criterion used to obtain criterion scores is important and no single criterion can fully represent effective teaching. The most often used criterion in teaching effectiveness research to correlate with student ratings is student achievement (Ory & Ryan, 2001). Ratings assigned to instructors by trained observers are another criterion measure. For example, the instructors who received high and low ratings from students also were found to teach differently when observed by trained observers (Murray, 1983).

Lastly, evidence based on relations to other variables is validity generalization. Analyzing this kind of evidence can indicate the degree to which this evidence can be generalized across new situations, especially when the student ratings are used to evaluate teaching effectiveness for different courses, time, or in different institutes. It is also important to show that a valid comparison can be made between ratings that are used for different purposes. The study of validity generalization can be complicated but provides useful information for the analysis of the dependability of student ratings. Research has examined the ability to generalize student ratings to different contexts. All findings consistently showed that student ratings collected from elective and required courses need to be interpreted differently (Brandenburg & Slinde, 1977; Costin et al., 1971; Feldman, 1978).

2.4.5 Collecting validity evidence based on survey consequences

How does evidence of the intended and unintended consequences of student ratings inform validity decisions and use of the survey? This source of validity evidence can support the implication inference and it is the latest validity evidence receiving attention in recent years. With so many institutions using student ratings in personnel decisions, the consequences of their use need to be addressed, including both positive and negative consequences (Ory & Ryan, 2001). The examination of survey consequences can be obtained from an investigation of the existence of the listed consequences, analysis of consequences, faculty personnel decisions, or classroom practices. The following unintended consequences occur when using student ratings (Crumbley, Flinn, & Reichelt, 2010; Ory & Ryan, 2001);

a. Instructors alter their teaching in order to receive high ratings,
b. Students reward poor teaching by believing they can give high ratings in return for high grades,
c. The institution rewards poor teaching, and
d. The rating process becomes a meaningless activity that is performed by students and instructors only because it is mandated.

However, this type of evidence was not a focus of this study.

In summary, when using student ratings of instruction to assess teaching effectiveness, there are validity concerns about the degree to which evidence and theory support the interpretations and uses of the instrument. This review develops an argument approach to validity for student ratings of instruction. Its intended interpretations and uses are clearly stated at the beginning. The IUA are needed to support interpretations and uses of assessment results, including four inference types: scoring, generalization, extrapolation, and implication. Each of these inferences is associated with various assumptions and then they are evaluated by the validity arguments. Claims are stated and validity evidence are collected to support those claims. Several sources of validity evidence together can provide clear answer to various issues concerning the validity of student evaluations of teaching. With respect to research on student evaluations of teaching, researchers encountered some issues/concerns regarding different sources of validity evidence as discussed previously.

Nevertheless, based on a synthesis of the research literature, the student evaluations of teaching in general are: (1) valid for its intended interpretations and uses but can be affected by various situations; (2) either unidimensional or multidimensional in terms of what they propose to measure; and (3) useful for both formative and summative purposes but widely used for formative evaluation.

The next section is an overview of EFA, CFA and ESEM approach for examining internal structure of the survey, which can be used to support a specific type of validity evidence based on internal structure.

2.5 EXPLORATORY STRUCTURAL EQUATION MODELING (ESEM)

2.5.1 EFA versus CFA

Exploratory and confirmatory factor analysis (EFA and CFA) have been at the heart of psychometric research and are widely used in the development and refinement of psychoeducational assessment instruments (Floyd & Widaman, 1995; Reise, Waller, & Comrey, 2000). While EFA provides a more realistic presentation of the data with the allowance of item cross-loadings, CFA includes many methodological advances than the former does (e.g., goodness-of-fit, estimation of different models, inclusion of factors, or correlated uniquenesses). In addition, CFA is performed when the researcher has a priori hypothesis about the internal structure of the instruments. Items using CFA as compared to EFA are only allowed to load on their main factors, while cross-loadings on the other factors are constrained to be zero. CFA specifies the number, meaning, associations, and pattern of free parameters in the factor loading matrix before a researcher analyzes the data. In EFA, the factors can be extracted from freely estimated cross-loadings without specifying the number and pattern of loadings between the observed variables and the latent factor variables (Marsh, Morin, Parker, & Kaur, 2014; Marsh et al., 2009). CFA may be useful in applying to substantively important questions based on a priori hypothesis about the multidimensionality of student surveys of instruction.

CFA structures are much more restrictive than EFA structures and not often available in practice. Given that items are rarely pure indicators of their corresponding constructs, they are fallible in nature, thus at least some degree of construct-relevant association can be expected between items and the non-target, yet conceptually-related constructs (Morin, Arens, & Marsh, 2016). In psychological measurement, nonzero cross-loadings are inherent but such restrictive constraints (i.e., items can only load on one factor) could inflate CFA factor correlations and lead to biased estimates (Asparouhov & Muthén, 2009; Marsh et al., 2009). A review of simulation studies (Asparouhov, Muthén, & Morin, 2015) showed that even small cross-loadings (as small as 0.100) should be explicitly taken into account, otherwise, parameter estimates could be inflated and thus biased. Because of this, CFA models often do not fit the data well and there is a tendency to rely on extensive model modification to find a well-fitting model (Asparouhov & Muthén, 2009).

For the reasons given, a new approach of exploratory structural equation modeling (ESEM) (Asparouhov & Muthén, 2009; Marsh et al., 2010; Marsh et al., 2009) is outlined below which is an integration of EFA, CFA, and SEM. This approach has the potential to resolve this dilemma and has wide applicability to psychometric research.

2.5.2 An introduction to ESEM

ESEM is a less restrictive measurement model to be used in tandem with the traditional CFA models. This offers a richer set of a priori model alternatives that can be subjected to a testing sequence (Asparouhov & Muthén, 2009). CFA and ESEM models differ in two main ways: the use of theory in model specification and the treatment of nontarget loadings. Within CFA, all parameters are specified by the researcher a prior and represent distinct hypotheses about the associations between both observed and latent constructs (Bollen & Pearl, 2012). Often, each indicator is loaded on only one factor, with all other possible loadings set to zero (see Figure 1, CFA model). In contrast, the only priori information required to run an ESEM model is the number of factors. All other parameters are freely estimated. That is, in the same manner as an EFA model, all factors are allowed to load on all indicators (see Figure 1, ESEM model) (Booth & Hughes, 2014).

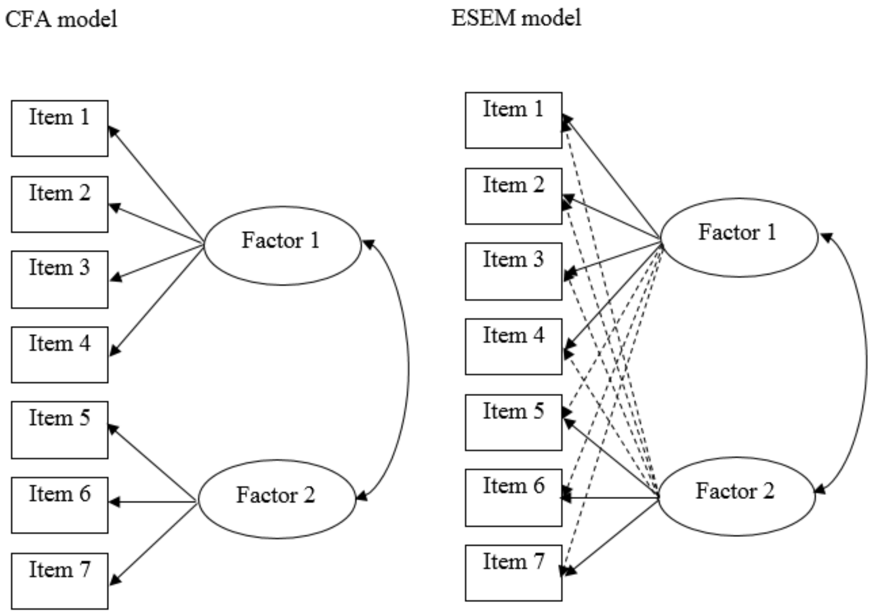


Figure 1. Simplified representations of the estimated CFA and ESEM models (a two-factor model)

The ESEM is viewed as a primarily confirmatory approach. It allows the analyst to control the expected factor structure. Within the ESEM framework, the researcher has access to typical SEM parameter estimates, standard errors, goodness-of-fit statistics, and statistical advances normally associated with CFA and SEMs (Asparouhov & Muthén, 2009). In a comprehensive review of ESEM (Marsh et al., 2014), most research studies that used ESEM has a clear a priori hypotheses regarding which indicators should load on which factors. That is, they used ESEM as a replacement for CFA in order to estimate cross-loadings. While, a few studies used ESEM as alternative to EFA in which the researchers had no clear hypothesis about the internal structure of the instruments.

The main purpose of this study was to develop and validate the student surveys of teaching effectiveness, represented by a set of manifest scores designed to reflect each of the multiple dimensions of teaching effectiveness. Hence, the use of ESEM as a viable confirmatory alternative to CFA can provide a potentially attractive alternative to the development of the student surveys with strong theoretical assumptions regarding the expected internal structure. Even when there is good support for the a priori factor structure on which the manifest factor scores are constructed, these scores are very limited in comparison with the reliance on latent constructs based on multiple indicators and taking into account measurement error (Marsh et al., 2009). In summary, to evaluate the multidimensional perspective of student surveys of teaching effectiveness, the ESEM can be used to provide better fit to the data because it integrates the flexibility of an EFA approach with the power of analyses that typically are conducted within a CFA framework.

2.5.3 Formal statistical basis of ESEM

Suppose that there are p dependent variables $Y = (Y_1, \dots, Y_p)$ and q independent variables $X = (X_1, \dots, X_q)$. Consider the general structural equation model with m latent variables $\eta = (\eta_1, \dots, \eta_m)$. The general ESEM model is described by the equations (Marsh et al., 2014).

$$Y = \nu + \Lambda\eta + KX + \varepsilon \qquad (2.5)$$

$$\eta = \alpha + B\eta + \Gamma X + \zeta \qquad (2.6)$$

Where Y is a dependent variable (p is a number of dependent variables)

X is an independent variable (q is a number of independent variables)

η is a vector of latent variables (m is a number of latent variables)

ν is a vector of intercepts

α is a vector of latent intercept

Λ is a factor loading matrix

B is a matrix of η on η regression coefficients

K is a matrix of Y on X regression coefficients

Γ is a matrix of η on X regression coefficients

ε is a vector of residuals for Y, and

ζ is a vector of residuals for η.

The standard assumption of this model is that the ε and ζ are normally distributed residuals with mean 0 and variance covariance matrix Θ and Ψ, respectively. Equation 2.5 represents the measurement model and equation 2.6 represents the latent variable model.

If the ESEM model includes a single factor ($m = 1$), then it is equivalent to the classic CFA model. When more than one factor is posited ($m > 1$), further constraints are required to achieve an identified solution (Asparouhov & Muthén, 2009). ESEM factors can be divided into blocks of factors so that a series of indicators is used to estimate all ESEM factors within a single block, and a different set of indicators is used to estimate another block of ESEM factors. However, specific items may be assigned to more than one set of ESEM or CFA factors. The assignment of items is usually determined on the basis of a priori theoretical expectations, on practical considerations, or perhaps post-hoc, based on preliminary tests conducted on the data (Marsh et al., 2014).

2.5.4 Estimation

All parameters in the ESEM model can be simply estimated by the maximum likelihood (ML) estimation, with weighted least square estimators, or with a robust alternative.

The estimation of the ESEM model consists of various steps. In the first step, an SEM model is estimated using the ML estimator. For each block of EFA factors the factor variance–

covariance matrix is specified as an identity matrix $(\Psi = \mathrm{I})$, giving $\frac{m(m+1)}{2}$ restrictions. The EFA loading matrix for the block (Λ) has all entries above the main diagonal (i.e., for the first m rows and column in the upper right corner of factor loading matrix, Λ), fixed to 0, providing remaining $\frac{m(m-1)}{2}$ identifying restrictions. This initial, unrotated model provides starting values that can be subsequently rotated into any other exploratory factor model with m factors. The asymptotic distribution of all parameter estimates in this starting value model is also obtained. Then, for each block of EFA factors, the ESEM variance covariance matrix is computed (based only on $\Lambda\Lambda' + \Theta$ and ignoring the remaining part of the model). In M*plus*, multiple random starting values are used in the estimation process to protect against nonconvergence and local minimums in the rotation algorithms. Although a wide variety of orthogonal and oblique rotation procedures are available (e.g., varimax, quartimin, geomin, target, equamax, parsimax, and oblimin), the choice of the alternative rotational procedures is still open for further study (Asparouhov & Muthén, 2009; Marsh et al., 2014; Marsh et al., 2009).

2.5.5 Goodness of fit

It is important to note that ESEM also requires researchers to evaluate how well the hypothesized models fit the data like other CFA/SEM models. A Chi-square (χ^2) test statistic is one test of model evaluation for the hypothesized model, examining whether the model implied covariance matrix is equal to the observed covariance matrix. There are some controversies regarding the use of χ^2 tests, specifically, it is less informative and moreover it is a function of sample size. Hence, it is not often of general interest (Clauser, Margolis, Holtman, Katsufrakis, & Hawkins, 2012).

In addition to the χ^2 test, a variety of fit indices to assess model fit are commonly reported in research articles. Most fit indices consider the model fit but also their simplicity (Clauser, Mazor, & Hambleton, 1994). χ^2 statistics and the goodness-of-fit indices are complimentary in nature. That is, some goodness-of-fit measures such as Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Tucker-Lewis Index (TLI) are a function of the chi-square and the degrees of freedom (Clauser, 1993). The population values of TLI and CFI range from 0 to 1, and values greater than .90 and .95 support acceptable and excellent model fits, respectively. The recommended cutoff values of less than .08 or .06 for RMSEA reflect reasonable and close fits to the data, respectively (Mazor, Clauser, & Hambleton, 1992).

In many studies, comparing the fit of alternative or competing models is useful. When any two models are nested, the set of parameters estimated in the more restrictive model is a subset of the parameters estimated in the less restrictive model. For purposes of model comparison, the relative fit of models testing fewer or more invariance constraints are more important than the absolute level of fit for any one model (Marsh et al., 2014). For example, Clauser, Nungester, Mazor, and Ripkey (1996) studied the relative fit of the ESEM model. The result suggests that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, there is reasonable support for the more parsimonious model. Chen (2007) also suggests that when the RMSEA increases by less than .015 there is support for the more constrained model. For indices that incorporate a penalty for lack of parsimony, such as the RMSEA and the TLI, it is possible for a more restrictive model to result in a better fit than a less restrictive model. However, these are all rough guidelines and should not be interpreted at "golden rules" (Mazor, Hambleton, & Clauser, 1998). To select the best model, a variety of different indices, professional judgment, a priori predictions, and common sense are required.

In summary, many psychological instruments have a well-defined factorial structure, but cannot be represented adequately within a CFA approach. CFA is usually too restrictive to provide acceptable goodness of fit in which each item is allowed to load on only one factor and all nontarget loadings are constrained to be zero. The misspecification of zero factor loadings systematically distorts the size of factor correlations. This can subsequently lead to distortions in structural relations. ESEM, an overarching integration of the best aspects of CFA, EFA, and SEM, provides a viable option. In many empirical studies, ESEM estimates of factor correlations are generally accurate. Hence, it can be used to provide validity evidence based on internal structure to support the interpretations and uses of the student survey of instruction.

2.6 DIFFERENTIAL ITEM FUNCTIONING

The examination of DIF in Likert- or rating-type scales has become increasingly important. There is an increased use of student surveys of teaching effectiveness for employment-related decisions that lead to major concern on an instructor's career stability. In developing a new student survey, it is essential to ensure that the data gained from the survey are valid for interpretations and uses of the survey. That is, the survey items need to be free from bias and not exhibit DIF against any certain group with particular characteristics that are not directly related to teaching effectiveness (e.g. gender, year in school, discipline).

The following sections are organized as follows: gender bias in student ratings of instruction, basic concept of differential item functioning (DIF) assessment, types of DIF, procedures for detecting DIF, and the proposed generalized Mantel-Haenszel (GMH) procedure for DIF assessment.

2.6.1 Gender bias in student ratings of instruction

The question of gender bias in student ratings of instruction has not been fully resolved. Earlier research provided conflicting results regarding the relationship between the gender of the student and student ratings of instruction. A number of research studies (Beran & Violato, 2005; McPherson & Jewell, 2007; McPherson, Jewell, & Kim, 2009) reported no differences between faculty ratings made by male and female students. Other studies reported differences in ratings given by male and female students. For example, the studies of Centra and Gaubatz (2000) and Kohn and Hatfield (2006) examined the impact of gender on student ratings of instruction and found that female students generally gave their teachers higher ratings than male students. Additionally, two studies by Santhanam and Hicks (2002) and Smith, Yoo, Farr, Salmon, and Miller (2007) also reported similar results of gender differences.

Therefore, to answer a question of "isn't there gender bias in student ratings?", differential item functioning (DIF) will need to be addressed to identify differences in the perception of specific items between male and female students.

2.6.2 Basic concept of DIF

DIF is a statistical characteristic used to describe items that have different measurement properties for two or more groups of comparable ability (Camilli & Shepard, 1994). In a survey or questionnaire with ordinal response scales, DIF occurs when different groups of people being matched by the intended survey traits endorse items in the survey to different degrees, regardless of item content. For example, an item classified as having DIF based on gender within a given survey, would indicate that at least one of the category responses may be more easily endorsed by males versus females given that they have the similar levels on the intended survey attribute (Weijters, 2006). Different probabilities of endorsing responses should contribute to differences in the intended-to-be-measured latent trait (e.g. attitude, perception, satisfaction) not group membership (e.g., gender, year of study, discipline). Thus, this item fails to capture participants' true attitudes or perceptions and could become a source of error in measurement (Camilli & Shepard, 1994).

DIF is a useful tool for developing new surveys or validating survey score inferences and can be performed as part of validity evidence. DIF and the extent of its presence is usually taken into account at the item level. The assumption behind DIF is that when participants are placed on the same metric using a matching variable, the probability of endorsing a certain response category in each item should be the same for every participant. Detecting DIF in ordinal items is complex because of the number of score levels and a test for DIF needs to be done at each level (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005).

Even though DIF can pose a major threat to the fairness and validity of psychometric measures, it does not necessarily guarantee that the item is biased (Angoff, 1993). This does not mean that DIF is unnecessary, but insufficient for analysis of item bias. An investigation using professional judgment can be conducted to evaluate plausible reasons for DIF. This enables researchers to correctly identify biased items and remove them from the survey (Clauser & Mazor, 1998).

2.6.3 Type of DIF

There are two types of DIF: uniform and nonuniform. Uniform DIF occurs when the probability of endorsing a category response for one group (e.g. males) is consistently higher than another (e.g. females) across the latent trait continuum. Whereas, nonuniform DIF signifies that the probability of endorsing a category response across the latent trait continuum between two or more groups is not the same. For example, the probability of endorsing a category response at the lower end of the latent trait continuum may be higher for one group than another (e.g. for males rather than females), but at the higher end of the latent trait continuum the probability is higher for the others (e.g. for females rather than males) (Swaminathan & Rogers, 1990).

2.6.4 Focal and reference group

DIF analysis is used to examine a possible bias that should be irrelevant to the intended trait in a measured survey. The probabilities of endorsing a response category of each item are compared between two groups of participants with similar overall skill levels on the trait. One group is called the reference group, while another is called the focal group. The focal group is the group of interest. The reference group is used as the standard to be compared with the focal group (Holland & Thayer, 1988). For this study, female and male is used to identify the focal and reference groups, respectively.

2.6.5 Matching variables

After identifying the focal and reference group, they both need to be matched based on their overall attitude on the trait being measured by the survey. This is done to ensure that both groups are comparable prior to the comparison of their survey item performance (Holland & Wainer, 1993). Two types of matching variables are internal and external. Usually, an observed sum of item scores (i.e., a total score) on a survey serves as an internal matching variable. Whereas, a participant's overall attitude on other surveys that measure a similar construct is used as an external matching variable (Clauser & Mazor, 1998). The matching score should be derived from DIF-free items plus the studied item in order to best control for Type I error (Su & Wang, 2005; Wang & Su, 2004; Zwick, Donoghue, & Grima, 1993).

2.6.6 Procedures for detecting DIF

There are two major kinds of DIF detection: a parametric approach, which assumes a specific item response model, and a nonparametric approach, which does not. Parametric approaches suffer from model misspecification because even a small amount of misfit may result in serious Type I error inflation (Bolt, 2002). Nonparametric approaches are powerful enough and do not require specific forms of item response functions, larger sample sizes, or intensive computation.

Various approaches have been developed to detect DIF for ordinal items. Empirical and simulation studies comparing the performance of DIF methods for items with ordinal response scales have reported that among all of the nonparametric approaches, the generalized Mantel-Haenszel (GMH) test (Mantel & Haenszel, 1959; Somes, 1986; Zwick et al., 1993) may be the best option for DIF detection.

There are several reasons why the GMH is a widely used nonparametric method. Several studies reported that the GMH had good Type I error rate control as well as high power for detecting uniform DIF. For instance, the GMH test was tested in three empirical studies (Spray & Miller, 1994; Tian, 1999; Zwick et al., 1993) and its Type I error was near 0.05 under most conditions. The GMH has shown good power (0.6 to 1.00) for detecting uniform DIF when sample sizes were moderately large (1200-2000) or when DIF magnitude was large (0.25 difference in b parameters). However, researchers reported mixed results in terms of its power for nonuniform DIF but Magis, Béland, Tuerlinckx, and de Boeck (2010) proposed a variation that reduces this limitation. When group ability differences existed, Tian (1999) found that the GMH showed higher power than other methods. Finally, three recent studies (Fidalgo & Madeira, 2008; Fidalgo & Scalon, 2010; Guilera, Gómez-Benito, & Hidalgo, 2009) concluded that the GMH procedure is conceptually simple, relatively easy to apply, offers a test of statistical significance, and provides an estimate of the effect size using the common odds ratio. Also, the GMH statistics can be calculated using easily accessible statistical software, including SPSS, OpenStat, GMHDIF, and LazStats.

2.6.7 Generalized Mantel-Haenszel (GMH) procedure

To detect DIF in ordinal items, a direct extension of the Mantel-Haenszel method: the generalized Mantel-Haenszel (GMH) procedure (Mantel & Haenszel, 1959; Somes, 1986; Zwick et al., 1993) is used in this study. The GMH procedure treats the response categories as nominal data. Controlling for a matching variable, the GMH procedure compares the reference and focal groups in the entire response distribution.

The GMH procedure discussed in this study is based on a contingency table framework (Zumbo & Hubley, 2003). To perform the GMH procedure, first, an observed total survey score is calculated to match an individual's overall trait levels with both the reference (r) and the focal (f) groups. Next, the data is arranged as a $2 \times T \times K$ contingency table, where T is the number of response categories in an ordinal item, and K is the number of matching variable levels (the survey score). The contingency table is constructed for each item at each observed survey score level. Hence, at each of the K levels, the data are organized into a $2 \times T$ contingency table and classified as belonging to the reference or focal group. The data structure for this contingency table is shown in Table 1.

Table 1. Data structure for $\boldsymbol{k^{th}}$ level of $\boldsymbol{2 \times T}$ contingency table

Group	Response variable categories					
	y_1	y_2	y_3		y_T	Total
Reference	n_{R1}	n_{R2}	n_{R3}	…	n_{RT}	n_{R+k}
Focal	n_{F1}	n_{F2}	n_{F3}	…	n_{FT}	n_{F+k}
Total	n_{+1}	n_{+2}	n_{+3}	…	n_{+T}	n_{++k}

The values y_1, y_2,…,y_T denote the T numbers of response categories in the ordinal item. The values n_{RTk} and n_{FTk} denote the numbers of the reference and focal group numbers, respectively, which receive a response category of y_t at k^{th} level of the matching variable. The "+" symbol denotes summation over a particular index.

The GMH method treats the response categories as nominal data. Following the notations in Table 1 yields

$$\mathbf{A}_k' = \left(n_{R1k}, n_{R2k}, \dots, n_{R(T-1)k}\right), \tag{2.7}$$

$$E(\mathbf{A}_k') = \frac{n_{R+k}\mathbf{n}_k'}{n_{++k}}, \tag{2.8}$$

$$\mathbf{n}_k' = \left(n_{+1k}, n_{+2k}, \dots, n_{+(T-1)k}\right), \tag{2.9}$$

$$V(A_k) = n_{R+k}n_{F+k}\left(\frac{n_{++k}diag(\mathbf{n}_k) - \mathbf{n}_k\mathbf{n}_k'}{n_{++k}^2(n_{++k}-1)}\right), \tag{2.10}$$

Where $diag(\mathbf{n}_k)$ is a $(T-1) \times (T-1)$ diagonal matrix with elements $\mathbf{A}_k'$; $E(\mathbf{A}_k')$ is a vector of length $(T-1)$; and $V(A_k)$ is a $(T-1) \times (T-1)$ covariance matrix. The GMH statistic may be expressed as

$$\chi_{GMH}^2 = [\sum \mathbf{A}_k - \sum E(\mathbf{A}_k)]'[\sum \boldsymbol{V}(\boldsymbol{A}_k)]^{-1}[\sum \mathbf{A}_k - \sum E(\mathbf{A}_k)]. \tag{2.11}$$

Under the null hypothesis of no DIF, the GMH statistic is distributed as $\chi^2(df = T-1)$. If the null hypothesis is rejected, DIF is found within a given item.

2.7 SUMMARY

Student surveys of teaching effectiveness are used for three purposes: (a) providing feedback to teachers for instructional improvement, (b) providing input for administrative decision making, and (c) providing evidence for demonstrating the performance of an institution as a part of internal quality assurance processes. It is crucial to ensure the technical quality of the student surveys of teaching effectiveness for these purposes.

The development of student surveys is followed by an established framework for validation. An ordinal scale is used to obtain the data from the student surveys. There are several measurement and statistics issues associated with using an ordinal scale. The issues related to reliability, validity, robustness, analyses, and number of categories are discussed in the literature review.

The interpretations and uses of scores from the student surveys of teaching effectiveness are validated based on the argument-based approach to validation. Four sources of validity evidence are accumulated, including validity evidence based on survey content, response processes and expert review of items, internal structure, and relations to other variables.

CTT is used to provide evidence for the student survey's technical quality, including the psychometric properties of the survey. The EFA, CFA and ESEM analyses can be used and compared to examine the internal structure of the survey. CFA is usually too restrictive to provide acceptable goodness of fit for most psychological instruments. ESEM integrates many of the advantages of CFA, SEM, and EFA and its flexibility is useful when applying it to the multiple dimensions of student ratings of teaching effectiveness.

Additionally, the survey items need to be free from bias and not exhibit DIF against any certain group with particular characteristics that are not directly related to teaching effectiveness such as gender, year in school, and discipline. It is crucial to examine DIF because the student surveys of teaching effectiveness can be used for employment-related decisions that impose serious consequences on an instructor's career stability. The generalized Mantel-Haenszel (GMH) procedure is employed to detect DIF in ordinal items. The GMH procedure treats the response categories as nominal data. Controlling for a matching variable (i.e., a total score), the GMH procedure compares the reference and focal groups in the entire response distribution.

In summary, the main purpose of this study was to develop and evaluate the psychometric quality and validity of student ratings of instruction for a new survey. There were four research questions;

Research question 1: To what extent does the content evidence support the construct definition?

Research question 2: To what extent do the relationships among survey items and survey components correspond to the construct dimension?

Research question 3: Is there gender differential item functioning in student ratings?

Research question 4: Are there relationships between student ratings and a similar measure of teaching quality and student achievement?

3.0 METHODOLOGY

In designing and developing a student survey intended to measure teaching effectiveness, the validity of the score interpretations and the psychometric quality of the instrument need to be established. The following is a discussion and overview of the various phases in accordance with professional psychometric standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) that apply to survey design and validation.

3.1 SURVEY DEVELOPMENT

Phase I: Determine the purpose of the survey and the construct to be measured by the survey.

First, the purpose and use of student ratings needed to be delineated. As mentioned in Chapter 1, Thai universities are preparing to adopt AUN-QA criteria in the development of their own instruments that measure student perceptions and reactions regarding teaching. In this manner, student ratings of instruction are used for three purposes: (a) providing feedback to teachers for instructional improvement, (b) providing input for administrative decision making, that is faculty promotion and tenure decisions, and (c) providing evidence for demonstrating the performance of an institution as a part of internal quality assurance processes (Galbraith et al., 2012; Kember et al., 2002; Richardson, 2005; Spooren et al., 2013).

Second, with respect to validity, it is important to define the construct of teaching quality. An extensive overview of research examining dimensionality of student ratings of instruction in higher education has revealed that the student rating survey should capture multiple aspects (dimensions) of teaching quality. Yet, there is no consensus regarding the number and the nature of these dimensions (see Spooren et al., 2013). That is, there are different numbers of dimensions captured in reported instruments of student ratings of instruction. Even though it is commonly accepted that teaching effectiveness is multidimensional, a few studies argue in favor of unidimensional construct (Apodaca & Grad, 2005; Oon, Spencer, & Kam, 2017). Furthermore, several studies reported an existence of unidimensional higher order factors that reflect general instructional skill (Burdsal & Harrison, 2008; Cheung, 2000; Harrison, Douglas, & Burdsal, 2004; Mortelmans & Spooren, 2009). However, both theory and empirical testing are necessary when it comes to dimensionality decision (Spooren et al., 2013).

Given that the pedagogical paradigms in today’s university teaching have multiple indicators of quality teaching along with the multifaceted aspects of AUN-QA criteria, this study aims to develop a survey of student ratings in a confirmatory manner. What to be measured in the student survey was defined prior to the development of the survey and derived from a thorough literature review of teaching dimensions that were consistently found to impact student learning in higher education. The research-based dimensions of good teaching were also aligned to the three conceptual-based criteria of AUN-QA that were employed to facilitate the assessment of teaching quality. The three criteria include criterion 4: teaching and learning approach, criterion 6: academic staff quality, and criterion 11: student satisfaction (output).

To identify and formulate important research-based instructional dimensions, a search of the recent literature on teaching effectiveness in higher education was conducted using

University Library Search Engine and Google Scholar^{TM}. The keywords such as “student ratings”, “student evaluation of teaching”, “teaching evaluation”, “teaching effectiveness”, and “teaching quality” were used to search for relevant studies. The criteria for study selection was year of publication (since 2000) and level of education (higher education) to ensure relevant and high-quality research. In sum, there was a final database of 30 relevant studies (26 journal articles, 2 book chapters, 1 doctoral dissertation, and 1 report) that discussed teaching dimensions that tend to predict student learning (Barnes et al., 2008; Barth, 2008; Basow & Montgomery, 2005; Chatterjee et al., 2009; Coffey & Gibbs, 2001; E. H. Cohen, 2005; Díaz, Swan, Ice, & Kupczynski, 2010; Feldman, 2007; Fresko & Nasser, 2001; Ginns, Prosser, & Barrie, 2007; Glazerman et al., 2011; Gursoy & Umbreit, 2005; Hativa, Barak, & Simhi, 2001; Keeley, Furr, & Buskist, 2010; Keeley, Smith, & Buskist, 2006; Kelly, Ponton, & Rovai, 2007; Kember & Leung, 2008; Marsh, 2007; Marsh et al., 2009; Mortelmans & Spooren, 2009; Renaud & Murray, 2005; T. R. M. Roberts, 2008; Safavi, Bakar, Tarmizi, & Alwi, 2012; Sedlmeier, 2006; Shevlin, Banyard, Davies, & Griffiths, 2000; Simendinger et al., 2017; Spooren, 2010; Spooren, Mortelmans, & Thijssen, 2012; Toland & De Ayala, 2005; Tucker, Oliver, & Gupta, 2013). The next step was to extract information regarding teaching characteristics/ dimensions/ domains/ factors that were found in the studies and determine how frequently each dimension was mentioned in all 30 relevant studies. An example of extracted information is shown in Table 2. For a complete list of all extracted information, please see Appendix A.

Table 2. Example of teaching effectiveness dimensions extracted from literature-based studies

Basow and Montgomery (2006) – six dimensions	Fresko and Nasser (2001) – eight dimensions	Tucker, Oliver, and Gupta (2013) – ten dimensions	Hativa, Barak, and Simhi (2001) – seven dimensions
Scholarship	Course content	Knowledgeable	Lesson organization
Organization/clarity	Course organization	Organized	Lesson clarity
Instructor – group interaction	Use of instructional aids	Encourages active student participation with learning	Making a lesson interesting/engaging
Instructor – individual interaction	Use of instructional strategies	Communicate clearly	Classroom climate
Dynamism/enthusiasm	Course assignments	Enthusiastic	Freedom to think
Overall effectiveness	Treatment of students	Approachable	Overall effectiveness
	Grading procedures	Sensitive to student learning	Classroom preparation
	Level of course demands	Available for help	
		Provides useful feedback	
		Helps student to learn	

Once the information regarding dimensions of teaching effectiveness was extracted, a synthesis table was constructed to determine how frequently each dimension was mentioned in all 30 relevant studies (see Table 3). The dimensions receiving the most support from the literature were of interest.

Table 3. Frequency tabulation of dimensions found in literature

Dimensions of teaching	Frequency mentioned in 30 relevant studies
Fairness of evaluation	22
Clarity of instruction	20
Preparation and organization	17
Knowledge of subject matter	13
Availability and helpfulness	12
Class interaction	12
Stimulation of interest in the course	10
Classroom management	10
Enthusiasm/interest in teaching	9
Intellectual expansiveness	9
Intellectual challenge/difficulty	9
Use of supplementary materials	9
Professional competency	8
Active learning	7
Grading timeliness	7
Teacher rapport	7
Workload	6

Dimensions of teaching	Frequency mentioned in 30 relevant studies
Overall instructor	6
Course value	5
Encouragement of independent thought	5
Sensitivity to student learning	4
Overall course	4
Creativity	3
Elocutionary skills	3
Ethic	3
Overall GPA	2
Respect for students	2
Impact of instruction	1
High expectation	1
Research	1
Individualization of teaching	1

The dimensions with the frequency of 6 and higher were targeted because they represent the dimensions receiving high support from the literature. The overall dimensions include fairness of evaluation, clarity of instruction, preparation and organization, knowledge of subject matter, availability and helpfulness, class interaction, stimulation of interest in the course, classroom management, enthusiasm/interest in teaching, intellectual expansiveness, intellectual challenge/difficulty, use of supplementary materials, professional competency, active learning, grading timeliness, teacher rapport, workload, and overall instructor rating.

The next task was to relate and select those dimensions that were most important to the three AUN-QA criteria. As a consequence, five dimensions including intellectual challenge, stimulation of interest in the course, active learning, class interaction, and intellectual expansiveness were chosen to measure criterion 4: teaching and learning approach. There was considerable overlap between the selected dimensions of teaching and criterion 4 of AUN-QA. For AUN-QA criterion 6: academic staff quality, there were also five dimensions including preparation and organization, knowledge of subject matter, classroom management, clarity of instruction, and use of supplementary materials that were chosen. Finally, four dimensions including professional competency, availability and helpfulness, grading timeliness, and fairness of evaluation were considered important as measures of criterion 11: student satisfaction (output).

An extensive review of the literature on student ratings of instruction (Spooren et al., 2013) indicated that institutions should be able to select the aspects that are most important, according to their educational vision and policy, when formulating a list of teaching dimensions. Bangkok University's vision is a creative and quality institution with a mission of producing graduates with virtue and independence. Therefore, the dimensions of encouragement of independent thought and creativity were also chosen and considered as important measures of criterion 4 and a dimension of ethics/morality was chosen to measure criterion 6 of AUN-QA even though those three dimensions have a small frequency found in the literature.

When a list of important dimensions was established from a commonality among the literature review and AUN-QA criteria, the selected dimension of teaching quality was then transformed into an item that asks students in terms of their perceptions and experiences on a range of teacher- and course-related aspects. The items that were consistent with those dimensions and criteria were selected and adapted from an existing catalog (Arreola, 2007) and the process of selecting/ adapting was done in phase II. Arreola's (2007) catalog consists of 525 items that may be used to develop a customized student rating survey in higher education. These items are offered as a beginning resource to aid the construction of a student rating survey. The items have been divided into four categories, including instructional design, instructional delivery, instructional assessment, and course management. Additionally, the items have also been listed in two additional categories: self-reported course impact on the student and alternate and supplementary teaching/learning environments (laboratory and discussion, clinical, seminars, team teaching, and field trip).

Phase II: Write or select items.

To write or select possible items, the important dimensions of effective teaching derived from the first step were reviewed in order to search for the shared characteristics with the AUN-QA criteria. The corresponding items relating to the student learning and teaching characteristics were then selected from Arreola's catalog (Arreola, 2007). The items in that catalog were designed to assess students' perceptions and reactions to teaching. They were described as low-inference items which require less judgment as an observer and therefore increases objectivity. For example, instead of asking "The instructor is clear", a comparative low-inference item is: "The instructor refers to experiences or examples to clarify concepts". This item is within the clarity of instruction dimension.

Overall, 24 items that appeared to capture each dimension and criterion were identified. Some of them were adopted, whereas some items were adapted to ensure their consistency with Bangkok University and AUN-QA criteria. Each specific item will be used to provide useful specific information for instructional improvement.

As discussed earlier the construction of this survey was guided by the confirmatory approach (i.e., it began with what good teaching is and the selection of the dimensions that represent effective teaching from the theory and research). The next important question to be considered was a way to obtain an overall quality of teaching. An indicator of the overall score can be derived from a summed score from several items of which it is known to measure a unidimensional construct (d'Apollonia & Abrami, 1997; Kolitch & Dean, 1999). This score can be used for summative purposes such as faculty promotion, tenure decision, and institution evaluation. However, the student survey was hypothesized to measure multidimensions, therefore, the scores on different dimensions should not be summed to represent overall teaching competency. Rather, it is recommended to present the results on individual items or specific dimensions when working on formative purposes (Spooren et al., 2013; Tucker et al., 2013). To fill the gap, a global (single) item asking students to express their judgments of the instructor's overall teaching effectiveness was added to provide a clear measure of overall teaching effectiveness. As suggested by Abrami (1985) and Cashin and Downey (1992), global items also accounted for a significant amount of the variance and can be comparably used with summed scores from a unidimensional aspect. Overall, there was a total of 25 items in an initial version of the student survey (see Table 4).

Table 4 displays a list of important dimensions that was established from the review of the literature and AUN-QA criteria with their associated frequencies and survey items. The items with an asterisk reflect Bangkok University’s vision and mission of creativity, virtue, and independence as aforementioned.

Table 4. Development of survey items and their associated dimensions and AUN-QA criteria

Dimension	Frequency	Corresponding survey item
AUN-QA criterion 4: teaching and learning approach		
Intellectual challenge	9	1. The instructor raises challenging questions and problems.
Stimulation of interest in the course	10	2. The activities in class keep me interested and motivated.
Active learning	7	3. The instructor helps to keep me engaged and participated in productive learning.
Class interaction	12	4. The instructor maintains a classroom atmosphere where I feel comfortable to express ideas and ask questions.
Intellectual expansiveness	9	5. The instructor encourages me to apply the knowledge created in this class to my work or other non-class related activities.
Encouragement of independent thought	5	6. The instructor encourages me to work and think independently. *(Bangkok University’s vision and mission)
Creativity	3	7. The instructor attempts to stimulate creativity. *(Bangkok University’s vision and mission)
AUN-QA criterion 6: academic staff quality		

Dimension	Frequency	Corresponding survey item
Preparation and organization	17	8. The instructor is well prepared for class.
Knowledge of subject matter	13	9. The instructor is competent in his/her knowledge of the subject.
Preparation and organization	17	10. The instructor presents the course content in an organized manner.
Classroom management	10	11. The instructor uses appropriate teaching methods which helps my learning.
Preparation and organization	17	12. The instructor clearly explains the course objectives in the beginning of class.
Clarity of instruction	20	13. The instructor explains the subject matter clearly.
Clarity of instruction	20	14. The instructor refers to experiences or examples to clarify concepts.
Knowledge of subject matter	13	15. The instructor increases or improves my understanding about the subject matter.
Use of supplementary materials	9	16. The instructor uses a variety of instructional media/technology in class when applicable.
Ethic/morality	3	17. The instructor adds information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline. *(Bangkok University's vision and mission)

AUN-QA criterion 11: student satisfaction (output)

Dimension	Frequency	Corresponding survey item
Professional competency	8	18. There is close agreement between the announced objectives of the course and what is actually taught.
Availability and helpfulness	12	19. The instructor provides useful feedback that helps me understand my strengths and weaknesses.
Grading timeliness	7	20. The instructor provides feedback in a timely fashion.
Availability and helpfulness	12	21. The instructor is reasonably accessible for help.
Fairness of evaluation	22	22. Assigned work is appropriate to course level and credits.
Fairness of evaluation	22	23. The exams reflect material emphasized in the course.
Fairness of evaluation	22	24. The instructor evaluates my work fairly.
Global item		
Overall instructor	6	25. Express your judgment of the instructor's overall teaching effectiveness.

After completing the initial item development procedure, a total of 24 items of teaching effectiveness were formed into seven dimensions based on the literature (Feldman, 2007). The dimensions were: (1) Planning and Preparation, including item 8, 10, 12, and 18; (2) Classroom Management and Environment, including item 1, 3, 10, 11, 16, and 17; (3) Knowledge of Subject Matter, including item 9 and 15; (4) Clarity of Presentation, including item 13 and 14; (5) Availability and Helpfulness, including item 19, 20, and 21; (6) Evaluation and Quality of Examination, including item 22, 23, and 24; (7) Student Outcome, including item 2, 5, 6, and 7.

In summary, phases I and II provided some of the validity evidence based on survey content because the construct dimensions of teaching quality were defined and investigated. All selected dimensions and items were consistent within the literature and AUN-QA criteria. Specifically, they were guided by the empirical research, important policy of the University, and pedagogical practice.

Phase III: Develop/select appropriate response scales.

An additional phase to be considered was the development of response scales for the items. The student rating survey uses a five-point rating scale (i.e. “to a very high degree”, “to a high degree”, “to a moderate degree”, “to a small degree”, and “hardly at all”) that asks students to rate 24 specific items about the extent to which those teaching behaviors occur. A qualitative six-point rating scale is used to rate the instructor’s overall teaching effectiveness. The rating scale categories include “very effective”, “effective”, “somewhat effective”, “somewhat ineffective”, “ineffective”, and “very ineffective”.

In an effort to make the items as objective as possible, Arreola (2007) suggested that the response scales mentioned earlier need to meet certain logical and technical requirements. First, the selected response scales need to be appropriate and logically follow the items. For example, the item “The instructor is accessible for help.” logically calls for a very high to very low degree response. Also, the item “Express your judgement of the instructor’s overall teaching effectiveness” logically calls for an effective or ineffective response. Second, the response scales must be parallel. The categories of “to a very high degree”, “to a high degree”, “to a moderate degree”, “to a small degree”, and “hardly at all” are the same type of response asking about the frequency. Whereas, the categories of “very effective”, “effective”, “somewhat effective”, "somewhat ineffective", "ineffective", and "very ineffective", are the same type of scale asking about quality. Third, the response scales must be balanced. Both selected response scales have equal numbers of positive and negative choices (e.g., two numbers of positive choices are "to a very high degree" and "to a high degree" and two numbers of negative choices are "to a small degree" and "hardly at all") to avoid skewness in the directions. Several studies suggested that the middle response such as "Neutral" or "Neither agree nor disagree" should not be used because some students will simply use it to avoid answering the items unless there is a specific meaning relative to the item (Arreola, 2007). However, a five-point rating scale is recommended for Quality Assurance system in Thailand (Clauser, Mazor, & Hambleton, 1991). Hence, the construction of middle response for this study is designed to have specific meanings ("to a moderate degree") and lies exactly in between two extremes of response scales. In this case, students are independent in selecting any responses in either directions in a balanced and symmetric way.

The global item uses a six-point rating scale because it is intended to elicit opinions of students concerning instructor's overall teaching effectiveness through their responses of "effective" or "ineffective" without including the choice of neutrality. Given that the global item will be used for summative purposes, increasing the number of response categories (i.e., from a five-point to six-point rating scale) enables enhanced discrimination as well as allows for students to respond to likely categories (Joshi et al., 2015; Wakita et al., 2012). Finally, every category on the scale in the student rating form is defined without using numbers or letters to denote them. This can help to ensure that each category is as objective as possible and results in reliable and valid data (Arreola, 2007).

Phase IV: Conduct field trials to gather the data needed for validity evidence determination.

To increase the rigor of validity evidence based on the survey content from Phases I and II, the 25 items were subjected to a review conducted by an external panel of the academic leaders, instructors, psychometricians, as well as students from Bangkok University, Thailand. They were purposefully selected in the study based on their backgrounds. The academic leaders were selected from different affairs (e.g. Academic Affair, Educational Innovation Affair, Administrative Affair, Financial Affair, etc.). Likewise, the selection of the instructors and students were taken from different schools and departments to represent different disciplines and ensure the suitableness for diverse teaching. The psychometricians were experienced in survey development and validation. There were 39 expert panelists (i.e. academic leaders, instructors, and psychometricians) and 42 key stakeholders (i.e. students) providing feedback on the survey.

They provided feedback on the dimensions, items, and response scales derived from the previous steps. This was done to determine whether the 25 survey items and 7 domains adequately covered the construct of teaching quality, as well as to determine the representative of the survey items of the teaching quality construct.

A questionnaire was developed by using content validity review criteria from previous content validity studies (Armstrong, Cohen, Eriksen, & Cleeland, 2005; Thrush et al., 2007; Wynd, Schmidt, & Schaefer, 2003). It was sent to all panelists to complete by e-mail (see the form in appendix B). All panelists were asked to:

1. Rate how well each item assesses the domain according to a five-point response scale (“Extremely well”, “Very well”, “Moderately well”, “Slightly well”, and “Not well at all”).

2. Rate how relevant each item assesses the teaching quality according to a five-point response scale (“A great deal”, “A lot”, “A moderate amount”, “A little”, and “None at all”).

3. Provide what modifications are needed to improve the clarity and meaningfulness of the items.

4. Rate the overall comprehensiveness of the 25 survey items in representing teaching quality according to a five-point response scale (“Extremely”, “Very”, “Moderately”, “Slightly”, and “Not at all”).

5. Provide additional survey item(s) within domains that should be added to improve the comprehensiveness of the survey.

6. Indicate how adequate the 7 domains are in covering teaching quality (“Yes” or “No”). If “No”, they need to provide additional domain(s) that should be added.

7. Rate how appropriate the selected response scale is for the 24 specific survey items according to a five-point response scale (“Excellent”, “Very good”, “Good”, “Fair”, and “Poor”).

First, data analyses were completed to answer question 2 (rate how relevant each item assesses the teaching quality according to a five-point response scale). The dataset from a total of 81 participants were quantitatively analyzed using the Lawshe’s content validity ratio (CVR) (Haynes, Richard, & Kubany, 1995). It refers to the degree to which panelists find overlap or commonality between each survey item and the examined content. As can be seen below, approximately 60% of the total panelists rated items in each dimension. For dimension 1, 50 panelists rated all items. For dimension 2, 47 panelists rated all items, for dimension 3, 47 panelists rated all items, for dimension 4, 45 panelists rated all items, for dimension 5, 47 panelists rated all items, for dimension 6, 48 panelists rated all items, and for dimension 7, 48 panelists rated all items. The panelists that were used for calculating the CVR for each dimension had to have rated each item for the dimension. The CVR for each survey item in Table 5 was calculated by the following formula:

$$CVR = \left[\frac{\left(E - \left(\frac{N}{2} \right) \right)}{\left(\frac{N}{2} \right)} \right]$$

where E = number of panelists rating the survey item is essential (i.e. a great deal or a lot)

and N = total number of panelists. The cut-off value of CVR for the study was

.50 (at $p = .05$) given that N = 50 for dimension 1

.47 (at $p = .05$) given that N = 47 for dimension 2

.47 (at $p = .05$) given that N = 47 for dimension 3

.45 (at $p = .05$) given that N = 45 for dimension 4

.47 (at $p = .05$) given that N = 47 for dimension 5

.48 (at $p = .05$) given that N = 48 for dimension 6

.48 (at $p = .05$) given that N = 48 for dimension 7

As seen in Table 5, most of the CVR values of survey items across the seven domains were higher than the cut-off value except for item 7 and 8 on dimension 2: Classroom Management and Environment.

Table 5. Lawshe's *CVR*

Item	E (indicating either a lot, a great deal)	CVR by item	CVR by dimension

Item	E (indicating either a lot, a great deal)	CVR by item	CVR by dimension
D1: *Planning and Preparation* (valid case 50)			.68
1. The instructor is well prepared for class.	43	.72	
2. The instructor clearly explains the course objectives in the beginning of class.	41	.64	
3. The instructor presents the course content in an organized manner.	42	.68	
4. There is close agreement between the announced objectives of the course and what is actually taught.	42	.68	
D2: Classroom Management and Environment (valid case 47)			.52
5. The instructor maintains a classroom atmosphere where I feel comfortable to express ideas and ask questions.	39	.66	
6. The instructor uses appropriate teaching methods which helps my learning.	40	.70	
7. The instructor uses a variety of instructional media/technology in class when applicable.	33	**.40**	
8. The instructor adds information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline.	33	**.40**	

Item	E (indicating either a lot, a great deal)	CVR by item	CVR by dimension
9. The instructor raises challenging questions and problems.	35	.49	
10. The instructor helps to keep me engaged and participated in productive learning.	35	.49	
D3: Knowledge of Subject Matter (valid case 47)			.64
11. The instructor is competent in his/her knowledge of subject.	39	.66	
12. The instructor increases or improves my understanding about subject matter.	38	.62	
D4: Clarity of Presentation (valid case 45)			.78
13. The instructor explains the subject matter clearly.	41	.82	
14. The instructor refers to experiences or examples to clarify concepts.	39	.73	
D5: Availability and Helpfulness (valid case 47)			.63

Item	E (indicating either a lot, a great deal)	CVR by item	CVR by dimension
15. The instructor provides useful feedback that help me understand my strengths and weaknesses.	38	.62	
16. The instructor is reasonably accessible for help.	38	.62	
17. The instructor provides feedback in a timely fashion.	39	.66	
D6: Evaluation and Quality of Examination (valid case 48)			.74
18. The instructor evaluates my work fairly.	42	.75	
19. The exams reflect material emphasized in the course.	42	.75	
20. Assigned work is appropriate to course level and credits.	41	.71	
D7: Student Outcome (valid case 48)			.63

Item	E (indicating either a lot, a great deal)	CVR by item	CVR by dimension
21. The activities in class keep me interested and motivated.	39	.63	
22. The instructor attempts to stimulate creativity.	38	.58	
23. The instructor encourages me to work and think independently.	40	.67	
24. The instructor encourages me to apply the knowledge created in this class to my work or other non-class related activities.	41	.71	
Global item (valid case 48)			.54
25. the instructor’s overall teaching effectiveness.	37	.54	

Based on the result, only 1 survey item (i.e. item 8) was deleted. Item 8 was initially selected based on Bangkok University vision and AUN-QA policy even though its associated dimension had a small frequency found in the literature. However, the panelists who were staff and students from Bangkok University did not agree so it was removed due to its absence of support by the research literature and a need from the University.

Data from the same group of panelists were also analyzed to answer question 1 (rate how well each item assesses the corresponding dimension). The results are shown in Table 6.

Table 6. Frequency and percentage of each category on how well each item assesses the dimension

Item	Extremely well (5)	Very well (4)	Moderately well (3)	Slightly well (2)	Not well at all (1)	Percentage of (5) and (4)
D1: *Planning and Preparation* (valid case 77)						
1. The instructor is well prepared for class.	38	34	5	0	0	93.51%
2. The instructor clearly explains the course objectives in the beginning of class.	35	32	8	2	0	87.01%
3. The instructor presents the course content in an organized manner.	34	33	10	0	0	87.01%
4. There is close agreement between the announced objectives of the course and what is actually taught.	37	33	7	0	0	90.91%
D2: Classroom Management and Environment (valid case 76)						
5. The instructor maintains a classroom atmosphere where I feel comfortable to express the ideas and ask questions.	33	34	9	0	0	88.16%
6. The instructor uses appropriate	30	34	10	2	0	84.21%

Item	Extremely well (5)	Very well (4)	Moderately well (3)	Slightly well (2)	Not well at all (1)	Percentage of (5) and (4)
teaching methods which helps my learning.						
7. The instructor uses a variety of instructional media/ technology in class when applicable.	28	34	11	3	0	81.58%
8. The instructor adds the information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline.	27	30	17	1	1	75.00%
9. The instructor raises challenging questions and problems.	25	39	11	1	0	84.21%
10. The instructor helps to keep me engaged and participated in productive learning.	35	30	11	0	0	85.53%
D3: Knowledge of Subject Matter (valid case 76)						
11. The instructor is competent in his/her knowledge of subject.	38	33	5	0	0	93.42%
12. The instructor increases or	33	33	9	1	0	86.84%

Item	Extremely well (5)	Very well (4)	Moderately well (3)	Slightly well (2)	Not well at all (1)	Percentage of (5) and (4)
improves my understanding about subject matter.						
D4: Clarity of Presentation (valid case 76)						
13. The instructor explains subject matter clearly.	35	36	5	0	0	93.42%
14. The instructor refers to experiences or examples to clarify concepts.	35	33	8	0	0	89.47%
D5: Availability and Helpfulness (valid case 76)						
15. The instructor provides useful feedback that help me understand my strengths and weaknesses.	30	38	8	0	0	89.47%
16. The instructor is reasonably accessible for help.	29	36	11	0	0	85.53%
17. The instructor provides feedback in a timely fashion.	32	36	7	0	1	89.47%
D6: Evaluation and Quality of Examination (valid case 76)						
18. The instructor evaluates my work fairly.	33	36	7	0	0	90.79%
19. The exams reflect material	36	34	6	0	0	92.11%

Item	Extremely well (5)	Very well (4)	Moderately well (3)	Slightly well (2)	Not well at all (1)	Percentage of (5) and (4)
emphasized in the course.						
20. Assigned work is appropriate to course level and credits.	35	34	6	1	0	90.79%
D7: Student Outcome (valid case 76)						
21. The activities in class keep me interested and motivated.	27	35	13	1	0	81.58%
22. The instructor attempts to stimulate creativity.	25	39	9	3	0	84.21%
23. The instructor encourages me to work and think independently.	31	37	8	0	0	89.47%
24. The instructor encourages me to apply the knowledge created in this class to my work or other non-class related activities.	29	37	10	0	0	86.84%
Global item (valid case 76)						
25. the instructor's overall teaching effectiveness.	30	33	10	2	1	82.89%

From Table 6, for Dimension 1; 93.51% of the 77 panelists who had used the five-point scale indicated that item 1 assessed the dimension. For items 2, 3, and 4, the percentages were 87.01%, 87.01%, and 90.91%, respectively. For Dimension 2; 88.16% of the 76 panelists indicated that item 5 assessed the dimension. For items 6, 7, 8, 9, and 10, the percentages were 84.21%, 881.58%, 75.00%, 84.21%, and 85.53%, respectively. Note that, the result of item 8 was consistent with the result from the previous investigation so it was justified to remove item 8 from the student rating survey. For Dimension 3; 93.42% of the 76 panelists indicated that item 11 assessed the dimension. For item 12, the percentage was 86.84%. For Dimension 4; 93.42% of the 76 panelists indicated that item 13 assessed the dimension. For item 14, the percentage was 89.47%. For Dimension 5; 89.47% of the 76 panelists indicated that item 15 assessed the dimension. For items 16 and 17, the percentages were 85.53%, and 89.47%, respectively. For Dimension 6; 90.79% of the 76 panelists indicated that item 18 assessed the dimension. For items 19 and 20, the percentages were 92.11% and 90.79%, respectively. For Dimension 7; 81.58% of the 76 panelists indicated that item 21 assessed the dimension. For item 22, 23, and 24, the percentages were 84.21%, 89.47%, and 86.84%, respectively. Finally, 82.89% of the 76 panelists indicated that item 25 (global item) assessed the dimension.

Additionally, 38.67% of 75 panelists who had used the five-point response scale to answer question 4 indicated that the 25 survey items were extremely comprehensive in representing teaching quality. For very and moderately comprehensive, the percentages were 52.00% and 9.33%, respectively. 97.33% of 75 panelists rated yes to the adequacy of the 7 domains in the coverage of teaching quality (question 6). Finally, 87.81% of 74 panelists rated the appropriateness of the selected response scale used for the survey items as very good to excellent (question 7).

Next, the qualitative comments were reviewed. Based on the feedback and advice from this panel, the dimensions of teaching quality were refined. Four dimensions of teaching quality were adopted instead of seven dimensions that were proposed earlier. The new framework of teaching quality was defined by the studies of Fink (2013) and Meyer et al. (2016). An overview of the new framework is provided below and the rationale to support the items or potential changes is provided in Table 7.

This framework was developed based on the review and the literature and consists of four dimensions. First, *Organization and Structure* refers to the extent to which the course is organized in meaningful ways around learning objectives. The learning objectives expose students to what they can expect from the course. A review of Hattie (2009) concluded that there was a direct link between good learning objectives and student achievement. An ability of an instructor to communicate and remain true to the learning objectives is also considered important for effective teaching. Additionally, an instructor is able to make clear presentations, give clear assignments, and provide clear due dates for the duration of the course. Nevertheless, this dimension does not include a higher level of organization such as an instructor's ability to link the class content to students' prior knowledge or real life (Fink, 2013; Meyer et al., 2016). Second, *Assessment and Feedback* is another important aspect in instructional design. After the development of learning objectives, an assessment that aligns with those objectives helps to facilitate student learning. Hence, this dimension mainly deals with the alignment of learning objectives and assessment. In addition, the frequency of assessment and timely feedback improve student learning. Hattie (2009) found that learning was impacted by frequent measurement but not as much as when it was complemented by feedback regarding student misunderstandings. Assessments can occur during a class period or at the end and should occur frequently enough to observe whether students attain the learning objectives. Lastly, timely feedback is required to help students reflecting their learning (Fink, 2013; Meyer et al., 2016).

Third, *Personal Interactions* are defined as instructor-student interactions. The quality and frequency of communication from the instructor to students as well as the instructor's sincerity, respect, and concern for students fall within this domain. Hattie (2009) claimed that the relationships between students and instructors enhance learning. That is, in a caring learning environment, the instructor's sensitivity in observing student reactions and the instructor's friendliness in establishing good rapport with students contribute to effective learning (Fink, 2013; Meyer et al., 2016).

Finally, *Academic Rigor* supports a student-centered teaching environment. It focuses on critical thinking, intellectual challenges, and deeper learning and understanding. This can enhance knowledge-centered lectures through a thoughtful and challenging learning environment to promote long-term learning. Further, instructional technology can help to develop higher levels of cognition and produce complex learning. Encouragement by an instructor to apply the course's content to students' prior knowledge, explain their reasoning by using problem-solving procedure, or creatively think are examples of intellectual challenges (Fink, 2013; Meyer et al., 2016).

Table 7. Comments from the panelists and their associated rationale

Comments from panelists	Rationale
Dimension 2	
- Add an item regarding the ability of instructor to encourage students to integrate the old knowledge to new knowledge.	- One original item has similar content: "The instructor encourages me to apply the knowledge created in this class to my work or other non-class related activities".

Comments from panelists	Rationale
- Add an item about the instructor technique to stimulate cooperative classroom environment.	- One original item has similar content: “The instructor helps to keep me engaged and participated in productive learning”.
- Item 6 “the instructor uses appropriate teaching methods which helps my learning”, item 7 “the instructor uses a variety of instructional media/technology in class when applicable”, and item 8 “the instructor adds the information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline” are more relevant to "teaching skill/philosophy". The classroom management sounds more like at administrative level, such as keep the classroom tidy, make the classroom a safe place, etc.	- Item 6, 7, and 8 were revised to ensure they are aligned with the framework of teaching quality dimensions (Fink, 2013; Meyer et al., 2016).
- Item 8 “the instructor adds information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline” is not relevant and important to the domain.	- Item 8 was deleted.
Dimension 3	
- Need more items such as "the teacher can address my questions relevant to the class content".	- One original item has similar content: “The instructor increases or improves my understanding of the subject matter”.

Comments from panelists	Rationale
	- Added an item that reflects what being suggested: “The instructor is genuinely interested in helping students learn”.
- Add an item regarding the ability of instructor to apply class material to real life.	- One original item has similar content: “The instructor encourages me to apply the knowledge created in this class to my work or other non-class related activities”.
- Add an item that asks if the instructor keeps lecture and material updated.	- This comment is negligible and does not related to the teaching framework.
- Item 11 “the instructor is competent in his/her knowledge of the subject” should be separated into 2 items including (1) the instructor is competent in his/her knowledge of the subject and (2) the instructor has good knowledge beyond the textbook.	- Item 11 encompasses the intent of the suggested item.
- Item 12 “the instructor increases or improves my understanding about subject matter” should be classified into teaching skill domain.	- The item was reclassified into “Organization and Structure” dimension.
Dimension 4	
- Need more items such as "the teacher would repeat the key points and summarize his/her lecture nicely".	- One original item has similar content: “The instructor explains subject matter clearly”.

Comments from panelists	Rationale
- Add an item about instructor vocal delivery.	- This comment is negligible and does not relate to the teaching framework.
- Add the item that asks if the instructor stimulates class discussions.	- One original item has similar content: “The instructor helps to keep me engaged and participated in productive learning”.
- Item 14 “the instructor refers to experiences or examples to clarify concepts” sounds like a "teaching skill".	- This item was deleted due to the irrelevance with the teaching framework.
Dimension 5	
- Add an item that asks if the instructor provides a variety of outside resources to help students with learning.	- Outside resources such as library and computer lab are already provided by the supporting academic department in the university to all students and they are evaluated every semester.
- Item 17 “the instructor provides feedback in a timely fashion” is not important since time should not be considered as Availability and Helpfulness.	- Timeliness of information reported to students is considered important for student learning (Meyer et al., 2016).
Dimension 6	
- May add some reverse coding items, such as "I found some assigned work/exams are too difficult to complete".	- It is not necessary because adding an item in the different direction can reduce a reliability (by interfering with the

Comments from panelists	Rationale
	correlation among items) and it does not really solve the problem of acquiescence bias (Krosnick & Presser, 2010).
- Add the item regarding the use of authentic assessment in class.	- Authentic assessment may not apply to all classes and has multiple meanings.
- Not only an exam can be used to evaluate student performance. Should combine several approaches and consider any errors that may arise.	- Added the terms of graded assignments, reports, etc. in addition to tests.
Dimension 7	
- I would describe outcome from academic and nonacademic.	- Items from Student Outcome were reclassified into both academic and nonacademic outcomes.
- May add some items such as "I feel my understanding of a subject was improved".	- One original item has similar content: “The instructor increases or improves my understanding about subject matter”.
Others	
- Better to have the same number of items in each domain to make a valid comparison.	- Reviewed the literature and found that to include an equal number of items from each domain would benefit from examining the dimensionality and local dependency of items (Baghaei, 2008;

Comments from panelists	Rationale
	Linacre, 2015). - Revised the number of survey items by adding the relevant items and deleting some irrelevant items, the revised version includes; Dimension 1: Organization and Structure = 6 items Dimension 2: Assessment and Feedback = 5 items Dimension 3: Personal Interactions = 5 items Dimension 4: Academic Rigor = 6 items
- Domain 3-6, I suggest adding more items. You may want at least 3-4 items in each domain. 3 items are considered as a minimum for identifying a latent factor.	- Combined domains and a minimum of 4 items in each domain.
- Adding at least one more domain as teaching skill/philosophy. But be careful, 7 or 8 domains are a lot, you may not end up finding these domains independent to each other later on in your analysis (i.e., EFA/CFA). You can keep on adding domains, but you may likely find some of	- Modified the domains based on the literature: 4 dimensions of teaching quality that were originally defined by the work of Fink (2013) and Meyer et al. (2016) were adopted instead of 7 domains including "Organization and Structure", "Assessment

Comments from panelists	Rationale
these domains loading one or two secondary factors.	and Feedback", "Personal Interactions", and "Academic Rigor".

The main purpose of the previous phases was to further acquire validity evidence based on the survey content. The results were used to revise and reduce the number of items. Based on the results from both quantitative and qualitative analyses, the survey items were then refined and shown in Table 8.

Table 8. The revised survey items

Revised Item	Dimension
1. The instructor is well prepared for class.	Organization and Structure
2. At the beginning of the course, the instructor specifies in detail course materials and grading procedures.	Organization and Structure
3. The instructor presents the course content in an organized manner.	Organization and Structure
4. There is close agreement between the announced objectives of the course and what is actually taught.	Organization and Structure
5. The instructor explains the subject matter clearly.	Organization and Structure
6. The instructor increases or improves my understanding of the subject matter.	Organization and Structure
7. The instructor creates a classroom atmosphere where I feel comfortable to express my ideas and ask questions.	Personal Interactions

Revised Item	Dimension
8. The instructor is interested in helping students learn.	Personal Interactions
9. The instructor is friendly.	Personal Interactions
10. The instructor keeps me engaged and helps me actively participate in productive learning.	Personal Interactions
11. The instructor is reasonably available for help.	Personal Interactions
12. The instructor provides useful feedback that helps me understand my strengths and weaknesses.	Assessment and Feedback
13. Graded assignments, tests, papers, homework, etc., are returned promptly.	Assessment and Feedback
14. Grading in the course is fair and consistent.	Assessment and Feedback
15. The exams reflect material emphasized in the course.	Assessment and Feedback
16. Assigned work is appropriate to course level and credits.	Assessment and Feedback
17. The course is intellectually stimulating.	Academic Rigor
18. The instructor uses a variety of instructional media/technology in class to enhance learning when applicable.	Academic Rigor
19. The instructor attempts to stimulate creativity.	Academic Rigor
20. The instructor encourages me to work and think independently.	Academic Rigor
21. The instructor encourages me to apply the knowledge learned in this class to my work or other non-class related activities.	Academic Rigor
22. The course encourages me to read further in the area.	Academic Rigor

Revised Item	Dimension
23. Express your judgement of the instructor’s overall teaching effectiveness.	Overall teaching effectiveness.

The final phase for item refinement was a pilot study. The purpose of the pilot study was to explore items and the internal structure in relation to the proposed four-dimension teaching model. This included an examination of the 23 items and a subsequent examination of the internal structure.

3.2 PILOT STUDY

3.2.1 Procedure

The redesigned survey was piloted with five undergraduate classes offered in the Summer term of 2017 at Bangkok University, Thailand. The students volunteered to be in the pilot study. All instructors from those 5 classes received a brief description of the study from the researcher, and the potential benefits and risks if they participated in the study. The link to the online pilot survey (both English and Thai version) was sent to all students via email during their final examination weeks and open for student responses for four weeks. They can choose which language they prefer. The instructors were asked to remind the students in their classes to complete the pilot survey online. The students from five classes were asked to respond to the student rating survey with several demographic items (i.e. student ID, gender, school, department, overall GPA, year of study). The student rating survey had 22 quantitative items with a five-point rating scale, 1 global item with a six-point rating scale, and 1 qualitative item asking students to provide additional comments. This open-ended question asked students to provide any suggestions that students think may contribute to the future quality and development of the class.

3.2.2 Instrument

The pilot student survey was created using Qualtrics thus the survey was provided to students online. It represented 4 different dimensions of teaching effectiveness. All 22 Likert items were used along with one global item in the online form. All 22 items were measured on a five-point response scale (i.e. “to a very high degree”, “to a high degree”, “to a moderate degree”, “to a small degree”, and “hardly at all”) and one global item was measured on six-point rating scale (i.e. “very effective”, “effective”, “somewhat effective”, “somewhat ineffective”, “ineffective”, and “very ineffective”).

3.2.3 Participants

With assistance of the instructors at Bangkok University and the Computer Center, there were 5 classes (9 sections) and a total of 109 students who participated in the pilot study. There were 83 female and 26 male students from 8 different schools (36.7% from School of Communication Arts, 35.8% from School of Humanities and Tourism Management, 8.3% from School of Digital Media and Cinematic Arts, 6.4% from School of Business Administration, 5.5% from School of Information Technology and Innovation, 5.5% from School of Accounting, 0.9% from School of Law, and 0.9% from School of Fine and Applied Arts) who participated in the pilot study. The mean student GPAX was 2.57 (*SD* = 1.11), ranging from 1.39 to 4.00. Almost half of the students were juniors (40.4%), 32.1% were seniors, 12.8% were freshmen, 11.9% were sophomores, and 2.8% were others.

They were selected purposely to represent a range of schools and departments. All 109 responses were submitted with completed response variables. The list of participants for each class is shown in Table 9.

Table 9. Number and percentage of participants in a pilot study

Class	Section	Instructor	#Participants	Total # Students	% of Class
AB339: English for In-flight Passenger Service	1011	Instructor 1	12	40	30.00%
	1012	Instructor 2	10	33	30.30%
CO301: Pre-cooperative Education	4811	Instructor 3	14	65	21.54%
	4812	Instructor 3	7	57	12.28%
	4813	Instructor 3	15	72	20.83%
	4814	Instructor 3	13	60	21.67%
EN001: Daily Conversation English	1353	Instructor 4	13	34	38.24%
GE111: Value of Graduates	2201	Instructor 5	13	95	13.68%
GE117: Mathematics for Daily Life	2401	Instructor 6	12	173	6.94%
Total students			109	629	17.33%

The percentages of students who completed the pilot survey in each class were low and varied from 6.94% to 38.24% with an overall response rate of 17.33%. Given that the survey was administered to all students during their final examination weeks, there was a possibility that they did not watch closely their emails and only focused on their examinations. Even though the survey was open for four weeks, it is commonly known that most of the students are away during the school break. Therefore, to overcome this limitation during a final period of data collection, the survey will be sent out during the last three weeks of teaching and open for a longer time. More reminder emails will be considered to enhance the response rate.

It is important to understand an effect of low response rate on the validity of the study. James, Schraw, and Kuch (2015) and Capa-Aydin (2016) indicated that low response rate increases a risk of using student rating scores for summative purposes. Precisely, data with less than 10 ratings should be interpreted with caution. It was recommended to collect data from at least two-thirds of the class to obtain representative student rating data. However, the purpose of the pilot study was not intended for summative evaluation thus this consequence was negligible.

To implement and interpret high-quality student ratings for both formative and summative purposes for the main study, collecting at least two-thirds of the class will help instructors and administrators to assess the quality of the student rating survey with valid information.

3.2.4 Data analysis

In the present study, the student rating survey was designed to measure multiple constructs of teaching effectiveness. Four dimensions of teaching quality were hypothesized based on the Fink (2013) and Meyer et al. (2016)'s teaching framework. The survey was intended to provide multiple pieces of information about teaching quality. For a pilot study, Classical Test Theory (CTT) was employed to examine psychometric properties of the student rating survey.

A total of 109 students participated in the pilot study and a total of 109 completed rating surveys were submitted by students. Even though the survey was administered in both English and Thai, only the Thai version was submitted to the Qualtrics system. The basic summary statistics (Mean and Standard Deviation) were examined to understand how student responses on the student rating survey were distributed (see Table 10). The polyserial correlations among 22 individual survey items and total score were also analyzed as a part of classical item analysis. These item-total correlations function the same as item discrimination. They were evaluated in terms of the extent to which there is a linear relationship between an item and its total score (the item-total correlations > .30 were used to guide assessment) (Gandek et al., 1998; Kidder & Judd, 1986). The corrected item-total polyserial correlation was used in the study instead of the uncorrected item-total polyserial correlation. In this case, the scores on an item and scores on the total survey were correlated when the item of interest was removed from the total criterion score. This was done because including the item of interest results in an inflated correlation and can be misleading (Wolf, 1967). The results are shown in Table 10.

Table 10. Summary statistics and item-total correlations for the pilot data

Item	Min	Max	*M*	*SD*	Polyserial item-total correlation
1. The instructor is well prepared for class.	1	5	4.23	.735	.938

Item	Min	Max	*M*	*SD*	Polyserial item-total correlation
2. At the beginning of the course, the instructor specifies in detail course materials and grading procedures.	1	5	4.28	.818	.911
3. The instructor presents the course content in an organized manner.	2	5	4.28	.818	.928
4. There is close agreement between the announced objectives of the course and what is actually taught.	2	5	4.23	.824	.913
5. The instructor explains the subject matter clearly.	1	5	4.25	.841	.925
6. The instructor increases or improves my understanding of the subject matter.	1	5	4.17	.811	.911
7. The instructor creates a classroom atmosphere where I feel comfortable to express my ideas and ask questions.	2	5	4.22	.809	.911
8. The instructor is interested in helping students learn.	1	5	4.26	.865	.959
9. The instructor is friendly.	1	5	4.30	.844	.934

Item	Min	Max	*M*	*SD*	Polyserial item-total correlation
10. The instructor keeps me engaged and helps me actively participate in productive learning.	2	5	4.27	.777	.954
11. The instructor is reasonably available for help.	2	5	4.13	.872	.932
12. The instructor provides useful feedback that helps me understand my strengths and weaknesses.	1	5	4.15	.815	.914
13. Graded assignments, tests, papers, homework, etc., are returned promptly.	1	5	4.15	.880	.926
14. Grading in the course is fair and consistent.	2	5	4.27	.777	.946
15. The exams reflect material emphasized in the course.	1	5	4.21	.806	.934
16. Assigned work is appropriate to course level and credits.	1	5	4.26	.787	.923
17. The course is intellectually stimulating.	1	5	4.26	.787	.870
18. The instructor uses a variety of instructional media/technology in	2	5	4.13	.818	.925

Item	Min	Max	*M*	*SD*	Polyserial item-total correlation
class to enhance learning when applicable.					
19. The instructor attempts to stimulate creativity.	1	5	4.15	.859	.936
20. The instructor encourages me to work and think independently.	2	5	4.28	.795	.935
21. The instructor encourages me to apply the knowledge learned in this class to my work or other non-class related activities.	2	5	4.34	.772	.936
22. The course encourages me to read further in the area.	1	5	4.17	.855	.853

All polyserial item-total correlations were high. This may or may not indicate that the items were measuring the same construct. With a correlation of more than .3, it indicated very good discrimination and none of the survey items were removed. However, the more each item correlates with the survey score as a whole, the higher all items correlate with each other. This most likely suggests that the survey is unidimensional for this sample of students.

Additionally, the means of the survey items ranged from 4.13 to 4.34 which are fairly similar. Inspection of the raw data indicated that some students were responding the same across items indicating they were not discriminating among response options.

Next, an analysis of the item-dimension (subscale) polyserial correlations were examined. This permitted evaluation of correlations in terms of the extent to which items correlated most strongly with other items measuring the same construct than they correlated with other constructs. The 22 survey items were hypothesized to represent four dimensions. The item-dimension correlations are presented in Table 11. The table shows correlations of items with hypothesized dimensions and with other dimensions.

Table 11. Item-dimension (subscale) correlations for hypothesized dimensions of a student rating survey (pilot data)

Item	Organization and Structure (OS)	Personal Interactions (PI)	Assessment and Feedback (AF)	Academic Rigor (AR)
1 (OS)	**.957***	.898	.911	.876
2 (OS)	**.917***	.872	.860	.862
3 (OS)	**.965***	.879	.899	.832
4 (OS)	**.954***	.855	.872	.833
5 (OS)	**.945***	.862	.883	.862
6 (OS)	**.961***	.842	.878	.831
7 (PI)	.868	**.931***	.877	.836
8 (PI)	.903	**.961***	.941	.894
9 (PI)	.858	**.951***	.897	.892
10 (PI)	.912	**.960***	.910	.906
11 (PI)	.853	**.957***	.883	.878
12 (AF)	.827	.880	**.929***	.890
13 (AF)	.902	.868	**.938***	.855

Item	Organization and Structure (OS)	Personal Interactions (PI)	Assessment and Feedback (AF)	Academic Rigor (AR)
14 (AF)	.908	.893	**.974***	.883
15 (AF)	.891	.905	**.950***	.874
16 (AF)	.882	.859	**.924***	.907
17 (AR)	.812	.796	.841	**.912***
18 (AR)	.851	.884	.902	**.922***
19 (AR)	.852	.875	.903	**.971***
20 (AR)	.865	.882	.891	**.965***
21 (AR)	.892	.881	.893	**.939***
22 (AR)	.775	.802	.831	**.889***

* Correlation of item with hypothesized dimension.

From the above table, the item-dimension correlation of each item showed the highest value with its hypothesized dimension. However, the correlations with other dimensions were still very high. Thus, the proposed multidimensional model of university teaching to the sample from the pilot study may not hold.

In terms of internal consistency, the Cronbach's α of all subscales were obtained: *organization and structure* (.948), *assessment and feedback* (.941), *personal interactions* (.929) and *academic rigor* (.938). The Cronbach's alpha showed that each subscale score reached high reliability.

Validity evidence based on relation to another variable was also obtained by examining the relationship between the results from the pilot data and a current survey used at Bangkok

University for Summer term. 109 students completed the pilot survey and scores ranged from 22 to 110 (*mean* = 92.807, *SD* = 18.230, *median* = 96). The distribution of survey scores was negatively skewed (skewness = -1.933). In terms of the summary statistics of the current survey used at Bangkok University, there were 13 survey items using 5-point Likert scale. 109 students were chosen to match the students who completed the pilot survey. Their scores ranged from 26 to 65 (*mean* = 49.514, *SD* = 10.578, *median* = 52). The distribution of survey scores was slightly negatively skewed (skewness = -.241). Due to the skewed distributions, a Spearman rank-order correlation was computed to assess the relationship between the proposed survey and the Bangkok University current survey used to measure teaching effectiveness. There was a significant strong positive relationship between two measures, r_s(107) = .725, *p* < .001.

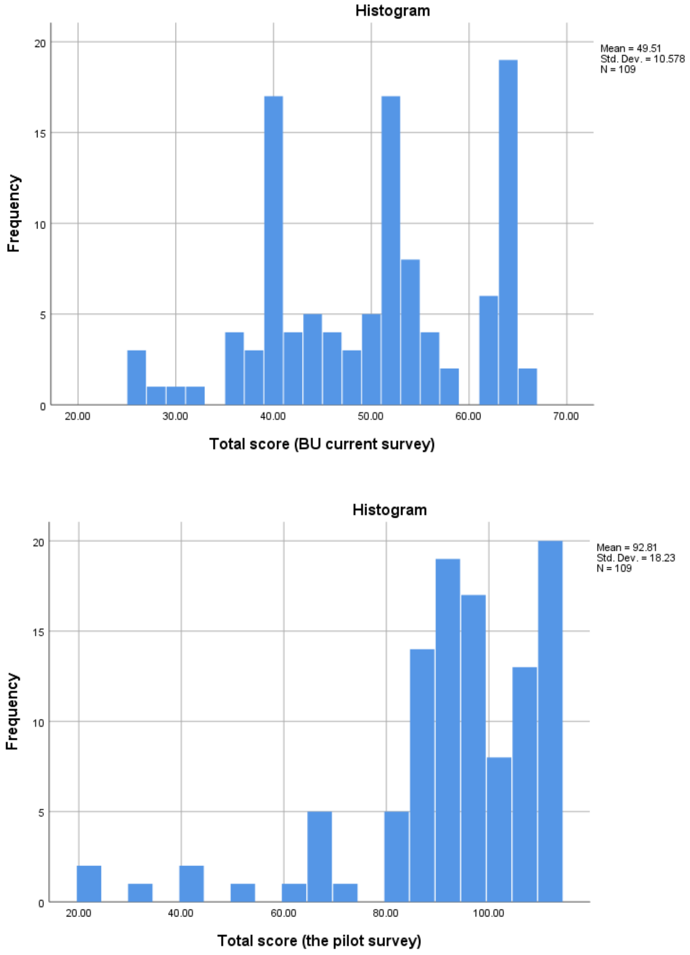


Figure 2. Frequency distribution of the total scores (the pilot survey and Bangkok University current survey)

Finally, the survey was administered to a large group of students during a full implementation study. It is assumed that the students will take the survey more seriously.

3.3 DATA COLLECTION FOR A FULL IMPLEMENTATION STUDY

The revised survey was administered to all undergraduate students who enrolled in 5 newly developed General Education (GenEd) classes at Bangkok University in the Spring semester of 2018. The courses were developed based on AUN-QA framework. The survey was administered in Thai only because all the students are Thais. The survey was administered by the instructors themselves, during normal class time, between the last three weeks of teaching of the Spring semester through Quatrics online survey system. Students were asked to complete the survey by clicking the survey link or scanning the QR code. Students were given the list of GE coursework units in which they were registered. Students were assured that the system was confidential and that instructors would have access to only average scores obtained from the students after all final grades were revealed. There were approximately 3,000 students (Bangkok University, 2018).

Five GenEd programs with 32 sections include:

GE001: Thinking Skills for Learning (section 1121 and 1131),

GE002: Citizenship and Social Dynamics (section 1241, 1251, 1261, and 1271),

GE003: Cultivating Entrepreneurial Mindset (section 1011, 1041, 1061, 1081, 1085, 1359, 1361, 1365, 1371, 1411, and 1021),

GE004: Technology and Innovation in the Future World (section 1121, 1131, 1359, 1411, 1061, 1085, and 1021), and

GE007: Art of Life (section 1241, 1251, 1261, 1271, 1391, 1401, 1451, and 1475)**.**

3.4 DATA ANALYSIS FOR A FULL IMPLEMENTATION STUDY

The data analyses for this study was completed to answer the following four research questions. Four types of validity evidence (evidence based on survey content, response processes and expert review of items, internal structure, and relations to other variables) were examined as well as demographic information were reported.

3.4.1 Data analysis for research question 1

Research question 1: To what extent does the content evidence support the construct definition?

3.4.1.1 Evidence based on survey content

The two preliminary studies in the pilot study provided evidence for this research question. The logical and empirical analyses of the adequacy of the survey content in representing the target dimension and of the relevance between the target dimension and the intended score interpretations and uses were evaluated in the two prior studies. This type of evidence was gathered from the well-designed survey development process including a thorough review of literature, an examination of the commonality between the literature and AUN-QA criteria, and experts' judgments in order to obtain a clearly defined construct of effective teaching. All survey items went through multiple rounds of expert evaluation and revision. Each category of the response scales was checked if it was applied by students. Finally, the survey consisted of four theoretical dimensions of teaching quality: (1) organization and structure, (2) assessment and feedback, (3) personal interactions, and (4) academic rigor.

3.4.2 Data analysis for research question 2

Research question 2: To what extent do the relationships among survey items and survey components correspond to the construct dimension?

3.4.2.1 Internal structure

During the pilot study, the correlations among dimension and within dimension were analyzed. That is, the item to total score polyserial correlations and the item to the dimension scores polyserial correlations were examined. The results from the pilot study did not support the proposed multidimensional model of university teaching.

To examine the internal structure of the student rating survey for a full implementation study, an exploratory factor analysis (EFA) was performed to offer initial, although tentative, insight into the internal structure. It can identify alternative grouping structures that could improve the functionality of the survey. Prior to the EFA analysis, the total sample data was investigated for univariate and multivariate normality. If the normality assumption is met, a series of analyses will be performed using maximum likelihood estimator (ML). In contrast, if the normality assumption is violated, a series of analyses will be performed using robust weighted least squares estimator (WLSMV), which provided standard errors and tests of model fit that are robust to the non-normality of the data. The MLSMV estimator is appropriate for ordered ordinal indicators (Brown, 2015). When interpreting the magnitude of the factor loadings, the guidelines of Comrey and Lee (1992) were applied: excellent above 0.71, very good between 0.63 and 0.70, good between 0.55 and 0.62, fair between 0.44 and 0.33, and poor below 0.32.

Model fit was assessed using a number of fit indices. The cutoffs for good or close fit (e.g., SRMR < .08, RMSEA < .06, TLI > .95, and CFI > .95) were used as well as cutoffs for acceptable levels of fit (e.g., CFI and TLI > .90). The model fit was assessed by rejecting models that did not meet bare minimum cutoffs for any one index and using combinatorial rules (i.e., TLI < .95 and SRMR > .06, CFI < .96 and SRMR > .06, or RMSEA > .06 and SRMR > .09) for rejection in cases where at least one index did not meet cutoffs for good or close fit (Hu & Bentler, 1999).

If multidimensionality was present, a confirmatory factor analysis (CFA) was used to confirm internal structure in a next step. CFA allows researchers to test a specified factor structure between items and dimensions, and allows for a rigorous test of facture structure in terms of fit with observed data. The fit indices were examined to see how well the multidimensional model fits the observed data. If there was not a good fit, exploratory structural equation modelling (ESEM) was performed later. ESEM combines the strengths of CFA and EFA within a SEM framework (Asparouhov & Muthén, 2009). ESEM allows for a complex structure where all items are permitted to load on all dimensions (see literature review for a more thorough discuss of EFA, CFA, and ESEM). In other words, if unidimensionality was truly present after performing EFA, the other two procedures (CFA and ESEM) were not used.

3.4.2.2 Descriptive statistics for subscales and total scores

Descriptive statistics were obtained for relevant subscales and total scores based upon extracted factors of the final survey. The basic summary statistics, including central tendency and variability, were examined to understand the distribution of survey scores for the entire sample.

3.4.2.3 Reliability

Finally, reliability estimates to examine the internal consistency of the survey scores were estimated using Cronbach's alpha coefficients. Coefficients were reported for all relevant subscales and total score based upon factors extracted from the factor analysis solution, section 3.4.2.1.

3.4.3 Data analysis for research question 3

Research question 3: Is there gender differential item functioning in student ratings?

3.4.3.1 Analysis of differential item functioning

The literature does not indicate a consistent difference in the student ratings made by males and females so a difference between student ratings made by male and female students was evaluated using Differential Item Functioning (DIF) analysis. DIF analysis was used to examine whether the survey item functions differently between male and female students. If so, it usually indicates individuals' membership affects their responses to a specific item in the survey, implying that item may be potentially biased against a subgroup.

Many DIF techniques have been developed for DIF detections, among which the Mantel-Haenszel (MH: Holland & Thayer, 1988; Mantel & Haenszel, 1959) is the most commonly used observed score and nonparametric technique. The original MH method can only conduct DIF analysis with dichotomous responses (e.g., 0 or 1). The extension of MH method, also called the generalized Mantel-Haenszel procedure and Mantel procedure (Agresti, 2013; Mantel, 1963) can be applied to both dichotomous and polytomous responses. For ordinal items, the choice between the generalized Mantel-Haenszel (GMH) procedure and the Mantel procedure basically depends on the pattern of DIF. The generalized Mantel-Haenszel (GMH) procedure (Mantel & Haenszel, 1959; Somes, 1986; Zwick et al., 1993) has good Type I error rate control as well as high power for detecting nonuniform DIF. It can detect DIF for more complex patterns compared to the Mantel procedure. Therefore, the GMH was employed to test the survey items for DIF. The presence of DIF items on a survey poses a threat to the validity of the interpretation and uses of survey scores.

The GMH procedure compares the entire response distribution between subgroups. It indicates the relation between the item and group membership, controlling for a matching score. From a result of the pilot study, it most likely suggested that the survey is unidimensional. Hence, a summed score that included the studied item was appropriate to be considered as a matching score. However, if the proposed multidimensional model of university teaching from the full implementation sample holds, matching on multiple subscale scores simultaneously is more superior to using the summed score.

First, descriptive statistics for female and male students were explored to see how students' responses distributed between subgroups. The summed scores (or the subscale scores) were computed and used as a matching score. The ICCs from the rating scale model (Andersen, 1977; Andrich, 1978) were used to examine the pattern of DIF by using the DIFAS program (Penfield, 2005). The type of DIF present was identified by examining the differences between the difficulty indices for each subgroup. These differences are labeled as "the conditional differences". If there is the same positive or negative direction across all the intervals of item performance, the DIF pattern is uniform. Conversely, the DIF pattern is non-uniform if the direction changes across the matching variable continuum.

Once the DIF pattern was confirmed, the GMH procedure was performed using the GMHDIF program (Fidalgo, 2011a, 2011b). The null hypothesis (H_0) specifies that there is no association between subgroups and item categories, controlling for the effect of the matching score, whereas the alternative hypothesis (H_1) indicates that the distribution of the response variable differs in nonspecific patterns across subgroups. The GMH chi square statistic is

$$\chi^2_{GMH} = [\sum_{k=0}^{K} \boldsymbol{N}_k - \sum_{k=0}^{K} E(\boldsymbol{N}_k)]'[\sum_{k=0}^{K} V(\boldsymbol{N}_k)]^{-1}[\sum_{k=0}^{K} \boldsymbol{N}_k - \sum_{k=0}^{K} E(\boldsymbol{N}_k)]. \qquad (2.12)$$

where k is a score on the matching criterion (k = 0, 1, 2, …, K), and $\boldsymbol{N}_k$ is a vector of T – 1 reference group frequencies (T = total number of categories). $E(\ )$ is the expected value under the null hypothesis of no DIF and $V(\ )$ is the variance. Under the null hypothesis of no DIF, the GMH statistic is distributed as $\chi^2(df = T - 1)$. Rejection of H_0 indicates that DIF found on the studied item is statistically significant.

As is always the case, statistical significance does not imply practical importance. DIF that is a small fraction of a score point in magnitude may not be worrisome even if it is statistically significant. In judging importance of DIF based on both statistical significance and the size of the DIF index, Holland and Thayer (1986) derived a descriptive measure of effect size of the MH procedure using a log-odds transformation of the odds-ratio into a difference on the delta scale called MH Delta-DIF. Educational Testing Service (ETS) uses this criterion to classify items as negligible DIF (“A”), moderate DIF (“B”), or large DIF (“C”). In developing a system for classifying polytomous items which is analogous to the ETS classification scheme for dichotomous items, I. M. Liu and Agresti (1996) proposed an estimator of the common odds ratio $(\hat{\alpha}_{LA})$ that is a natural generalization of the Mantel-Haenzel common odds ratio $(\hat{\alpha}_{MH})$ used for dichotomous items. To classify items into three categories, the significance of the GMH chi-square statistic was considered as well as the absolute value of $\hat{\alpha}_{LA}$. Using this adjustment to the original ETS scheme, a parallel scheme for classifying DIF in polytomous items as negligible (“A”), moderate (“B”), and large (“C”) is given by (Penfield, 2007; Penfield & Algina, 2003):

("A") if χ^2_{GMH} is significantly different from zero and $|\hat{\alpha}_{LA}| < 0.43$

("B") if χ^2_{GMH} is significantly different from zero and $|\hat{\alpha}_{LA}| \geq 0.43$, and either: $|\hat{\alpha}_{LA}| <$ 0.64, or $|\hat{\alpha}_{LA}|$ is not significantly greater than 0.43

("C") if χ^2_{GMH} is significantly different from zero and $|\hat{\alpha}_{LA}| \geq 0.64$, or $\hat{\alpha}_{LA}$ is significantly greater than 0.43.

The DIFAS program (Penfield, 2005) was used to calculate an estimator of the common odds ratio $(\hat{\alpha}_{LA})$.

With regard to the software programs used to answer this research question, the DIFAS program was used to examine the DIF pattern and calculate an estimator of the common odds ratio $(\hat{\alpha}_{LA})$. However, it cannot calculate GMH statistics (Penfield, 2005). Thus, the GMHDIF was used to calculate GMH statistics (Fidalgo, 2011a, 2011b). Both programs are free of charge and can be obtained by contacting the authors. Data entry runs in Windows which is user friendly. Both programs can be used for evaluating DIF in both dichotomous and polytomous items.

The GMHDIF program applies the purification procedure of two-stage DIF analyses in calculating GMH statistics. If there are any items found to exhibit DIF in the first stage, they are removed for calculating the matching criteria in the second stage. This program was chosen because it is easy to use compared to SAS software (SAS Institute Inc., NC). It automatically computes the total survey score and excludes those levels of the total score that only has one examinee. The GMHDIF program at this point is the most complete and simple program for detecting DIF using GMH statistics (Fidalgo, 2011b; Padilla, Hidalgo, Benitez, & Gomez-Benito, 2012; Penfield, 2005).

3.4.4 Data analysis for research question 4

Research question 4: Are there relationships between student ratings and a similar measure of teaching quality and student achievement?

3.4.4.1 Evidence based on relations to other variables

The fourth research question was analyzed through correlation analyses. The relationship between student ratings and another measure of instructional quality (i.e., result from the previous student rating form) was obtained in the pilot study. The correlation indicated a strong positive relationship. The relationships between student ratings and a measure of student achievement (i.e., overall GPA) was obtained in the full implementation study for the entire sample. It was expected that the student ratings and students' overall GPA would show a positive correlation because of a series of articles by Cohen (1980, 1981, 1982). He synthesized validity studies examining the relationship between overall student ratings and student achievement across different courses and instructors. Based on 68 multisection studies, Cohen found that the average correlations of overall instructor and overall course with student achievement were .43 and .47, respectively. The correlation coefficients were positive with a range between .01 to .90 with an exception of eight studies that had negative correlations (r = -.80, -.75, -.28, -.15, -.15, -.11, -.11, and -.04). He noted that the overall relationship between student rating and student achievement varied due to course levels and sample sizes. The magnitude of correlation coefficients decreased when the sample size was larger. Nonetheless, a positive relationship was generally expected to assure that student ratings reflect an impact of instructor/course on students to some extent. Additionally, these extended relationships can provide external validity evidence and are helpful in enhancing the meaning of score interpretations and uses.

3.4.5 Software of analysis

Data were first cleaned and all basic summary statistics (i.e. descriptive statistics, correlation, reliability coefficients, etc.) were explored using SPSS 23.0 software (IBM Corp., 2015). Exploratory factor analysis (EFA) was performed using Mplus (Muthén & Muthén, 1998-2012). If multidimensionality was present, confirmatory factor analysis (CFA) and exploratory structural equation modeling (ESEM) were proposed using Mplus (Muthén & Muthén, 1998-2012). For differential item functioning (DIF), the generalized Mantel-Haenzel chi-square statistic was estimated with GMHDIF program (Fidalgo, 2011a, 2011b). Finally, the DIFAS program (Penfield, 2005) was used to examine DIF pattern and calculate an estimator of the Liu-Agresti cumulative common log-odds ratio ($\hat{\alpha}_{LA}$).

4.0 RESULTS

This chapter provides results from both the pilot study and the full implementation study to answer the following four research questions:

1. To what extent does the content evidence support the construct definition?

2. To what extent do the relationships among survey items and survey components correspond to the construct dimension?

3. Is there gender differential item functioning in student ratings?

4. Are there relationships between student ratings and a similar measure of teaching quality and student achievement?

The validity evidence used to support research question 1 was obtained during two preliminary studies (i.e., the logical and empirical analyses) in the survey development phase and the pilot study. Similarly, a part of research question 4 was explored during the pilot study by correlating the results from the pilot data and a current survey used at Bangkok University to support external validity evidence. The results are reported in chapter 3 (methodology part) and summarized in the result section.

A correlation analysis was examined between scores from a student rating survey and students' overall GPA as a measure of student achievement in the full implementation study to answer research question 4. Research questions 2 and 3 were also studied during the full implementation study to obtain validity evidence to support the proposed interpretations and uses of student rating scores.

The results that answer all research questions are reported in this chapter.

4.1 DATA

4.1.1 Data screening

Data were collected from 5 newly developed General Education (GenEd) courses in compliance with AUNQA framework. 3,977 students were enrolled in 5 GenEd courses with 32 sections in the Spring semester of 2018. With assistance of the instructors at Bangkok University, there were a total of 2,256 students who participated in the full implementation study. Out of 2,256 total participants, 2,234 participants completed the entire survey. Twenty-two participants only completed some of the demographic items (i.e., subject, section, gender, school, and GPAX) and skipped or missed all 23 survey items. They were excluded from all of the analyses.

The list of participants for each course/section is shown in Table 12.

Table 12. Number and percentage of participants in a full implementation study

Course	Section	Instructor	#Participants	Total # Students	% of Class
GE001: Thinking Skills for	1121	Instructor 1	35	126	27.78
Learning	1131	Instructor 1	56	150	37.33
GE002: Citizenship and	1241	Instructor 2	34	78	43.59

Course	Section	Instructor	#Participants	Total # Students	% of Class
Social Dynamics	1251	Instructor 2	49	138	35.51
	1261	Instructor 2	72	137	52.55
	1271	Instructor 2	44	113	38.94
GE003: Cultivating Entrepreneurial Mindset	1011	Instructor 3	119	127	93.70
	1021	Instructor 6	92	161	57.14
	1041	Instructor 4	100	154	64.94
	1061	Instructor 5	77	86	89.53
	1081	Instructor 4	108	136	79.41
	1085	Instructor 5	94	102	92.16%
	1359	Instructor 3	102	123	82.93
	1361	Instructor 3	68	102	66.67
	1365	Instructor 4	135	160	84.38
	1371	Instructor 4	115	148	77.70
	1411	Instructor 3	82	108	75.93
GE004: Technology and Innovation in the Future World	1021	Instructor 6	76	157	48.41
	1061	Instructor 5	61	87	70.11
	1085	Instructor 5	68	94	72.34
	1121	Instructor 1	66	106	62.26
	1131	Instructor 1	89	150	59.33
	1359	Instructor 3	41	132	31.06
	1411	Instructor 3	75	110	68.18

Course	Section	Instructor	#Participants	Total # Students	% of Class
GE007: Art of Life	1241	Instructor 2	29	97	29.90
	1251	Instructor 2	38	137	27.74
	1261	Instructor 2	51	120	42.50
	1271	Instructor 2	33	114	28.95
	1391	Instructor 7	77	152	50.66
	1401	Instructor 7	53	107	49.53
	1451	Instructor 7	49	131	37.40
	1475	Instructor 7	46	134	34.33
Total students			2,234	3,977	56.17

The percentages of students who completed the survey in each class varied from 27.74% to 93.70% with an overall response rate of 56.17%.

4.1.2 Participant characteristics

SPSS 23.0 software (IBM Corp., 2015) was used to describe demographic characteristics of the 2,234 students. Of students taking the student rating survey, 4.07% were enrolled in GE001: Thinking Skills for Learning (91 students), 8.91% were enrolled in GE002: Citizenship and Social Dynamics (199 students), 48.88% were enrolled in GE003: Cultivating Entrepreneurial Mindset (1,092 students), 21.31% were enrolled in GE004: Technology and Innovation in the Future World (476 students), and 16.83% were enrolled in GE007: Art of Life (376 students).

The average student GPAX was 2.96 (*SD* = .761), ranging from 0.37 to 4.00. A further breakdown of students by gender and school is displayed in Table 13.

Table 13. Descriptive statistics of the demographic variables in the full implementation study (*n* = 2,234)

Variable		Number of Students	% of Students
Gender	Male	878	39.3
	Female	1356	60.7
School	School of Humanities and Tourism Management	687	30.75
	School of Business Administration	631	28.25
	School of Digital Media and Cinematic Arts	338	15.13
	School of Accounting	279	12.49
	School of Communication Arts	177	7.92
	School of Information Technology and Innovation	95	4.25
	School of Engineering	10	0.45
	School of Law	7	0.31
	School of Fine and Applied Arts	4	0.18
	School of Architecture	3	0.13
	School of Economics	3	0.13

There were more female than male students. The majority of students were from the School of Humanities and Tourism Management and the School of Business Administration, which accounted for 56.8% of the participants.

4.1.3 Item level descriptive statistics

Classical Test Theory (CTT) was employed to examine descriptive statistics (i.e., frequency, percentage, mean, and standard deviation) of the survey items in each dimension/subscale to understand how student responses on the student rating survey were distributed. Dimension 2; Personal Interactions, had the highest rating (*M* = 4.10, *SD* = .906), followed by dimension 4; Academic Rigor (*M* = 4.08, *SD* = .902), dimension 1; Organization and Structure (*M* = 4.07, *SD* = .906), and dimension 3; Assessment and Feedback (*M* = 4.03, *SD* = .912) (see Table 14). The means and standard deviations however were very similar across dimensions.

Most of students rated the teaching/learning behaviors described in each survey item occurred, “to a high degree” or “to a very high degree”. The items rated as “to a very high degree” most frequently were item 9; the instructor is friendly and establishes good rapport with students, 45.8%., item 7; the instructor creates a classroom atmosphere where I feel comfortable to express my ideas and ask questions, 43.6%, and item 20; the instructor encourages me to work and think independently, 43.5% (see Table 14).

Table 14. Item level frequencies and summary statistics for responses on the student rating survey (*n* = 2,234)

Item	Hardly at all n (%)	To a small degree n (%)	To a moderate degree n (%)	To a high degree n (%)	To a very high degree n (%)	*M (SD)*
Dimension 1: Organization and Structure (*M* = 4.07, *SD* = .906)						
1. The instructor is well	93	48	264	873	956	4.14
prepared for class.	(4.2)	(2.1)	(11.8)	(39.1)	(42.8)	(.993)
2. At the beginning of the	76	80	313	945	820	4.05
course, the instructor specifies	(3.4)	(3.6)	(14.0)	(42.3)	(36.7)	(.977)
in detail course materials and						
grading procedures.						
3. The instructor presents the	79	70	329	940	816	4.05
course content in an	(3.5)	(3.1)	(14.7)	(42.1)	(36.5)	(.977)
organized manner.						
4. There is close agreement	78	75	322	904	855	4.07
between the announced	(3.5)	(3.4)	(14.4)	(40.5)	(38.3)	(.986)
objectives of the course and						
what is actually taught.						
5. The instructor explains the	86	77	305	914	852	4.06
subject matter clearly.	(3.8)	(3.4)	(13.7)	(40.9)	(38.1)	(1.000)
6. The instructor increases or	88	76	349	898	823	4.03
improves my understanding	(3.9)	(3.4)	(15.6)	(40.2)	(36.8)	(1.007)

Item	Hardly at all n (%)	To a small degree n (%)	To a moderate degree n (%)	To a high degree n (%)	To a very high degree n (%)	*M (SD)*
of the subject matter.						
Dimension 2: Personal Interactions (*M* = 4.10, *SD* = .906)						
7. The instructor creates a classroom atmosphere where I feel comfortable to express my ideas and ask questions.	85 (3.8)	67 (3.0)	289 (12.9)	819 (36.7)	974 (43.6)	4.13 (1.005)
8. The instructor is interested in helping students learn.	81 (3.6)	74 (3.3)	312 (14.0)	894 (40.0)	873 (39.1)	4.08 (.992)
9. The instructor is friendly and establishes good rapport with students.	82 (3.7)	67 (3.0)	238 (10.7)	823 (36.8)	1024 (45.8)	4.18 (.991)
10. The instructor keeps me engaged and helps me actively participate in productive learning.	81 (3.6)	68 (3.0)	312 (14.0)	884 (39.6)	889 (39.8)	4.09 (.989)
11. The instructor is reasonably available for help.	73 (3.3)	97 (4.3)	359 (16.1)	918 (41.1)	787 (35.2)	4.01 (.991)
Dimension 3: Assessment and Feedback (*M* = 4.03, *SD* = .912)						
12. The instructor provides	76	77	338	921	822	4.05

Item	Hardly at all n (%)	To a small degree n (%)	To a moderate degree n (%)	To a high degree n (%)	To a very high degree n (%)	*M (SD)*
useful feedback that helps me understand my strengths and weaknesses.	(3.4)	(3.4)	(15.1)	(41.2)	(36.8)	(.981)
13. Graded assignments, tests, papers, homework, etc., are returned promptly.	78	86	361	914	795	4.01
	(3.5)	(3.8)	(16.2)	(40.9)	(35.6)	(.993)
14. Grading in the course is fair and consistent.	78	82	313	945	816	4.05
	(3.5)	(3.7)	(14.0)	(42.3)	(36.5)	(.982)
15. The exams reflect material emphasized in the course.	87	81	328	926	812	4.03
	(3.9)	(3.6)	(14.7)	(41.5)	(36.3)	(1.003)
16. Assigned work is appropriate to course level and credits.	83	89	309	966	787	4.02
	(3.7)	(4.0)	(13.8)	(43.2)	(35.2)	(.992)
Dimension 4: Academic Rigor (*M* = 4.08, *SD* = .902)						
17. The course is intellectually stimulating.	73	89	304	945	823	4.05
	(3.3)	(4.0)	(13.6)	(42.3)	(36.8)	(.977)
18. The instructor uses a variety of instructional media/technology in class to	77	80	299	892	886	4.09
	(3.4)	(3.6)	(13.4)	(39.9)	(39.7)	(.988)

Item	Hardly at all n (%)	To a small degree n (%)	To a moderate degree n (%)	To a high degree n (%)	To a very high degree n (%)	*M (SD)*
enhance learning when						
applicable.						
19. The instructor attempts to	81	75	266	893	919	4.12
stimulate creativity.	(3.6)	(3.4)	(11.9)	(40.0)	(41.1)	(.989)
20. The instructor encourages	78	65	270	849	972	4.15
me to work and think	(3.5)	(2.9)	(12.1)	(38.0)	(43.5)	(.982)
independently.						
21. The instructor encourages	74	80	312	928	840	4.07
me to apply the knowledge	(3.3)	(3.6)	(14.0)	(41.5)	(37.6)	(.976)
learned in this class to my						
work or other non-class						
related activities.						
22. The course encourages me	88	75	323	945	803	4.03
to read further in the area.	(3.9)	(3.4)	(14.5)	(42.3)	(35.9)	(.996)

For the global item that is intended to elicit opinions of students concerning instructor’s overall teaching effectiveness, 91.3% of students rated instructor’s overall teaching effectiveness as, “somewhat effective”, “effective” or “very effective”. Table 15 displays descriptive statistics for the global item.

Table 15. Descriptive statistics for the global item (*n* = 2,234)

Item	Very ineffective n (%)	Ineffective n (%)	Somewhat ineffective n (%)	Somewhat effective n (%)	Effective n (%)	Very effective n (%)
Overall, I would rate this instructor	12 (.5)	68 (3.0)	110 (4.9)	394 (17.6)	854 (38.2)	796 (35.6)
Mean = 4.97, *standard deviation* = 1.044						

This global item could provide a clear measure of overall teaching effectiveness if a multidimensional model was confirmed. Scores from this item can be used for summative purposes such as faculty promotion, tenure decision, and institution evaluation.

4.2 CONTENT DOMAIN AND DEVELOPMENT

To answer research question 1: to what extent does the content evidence support the construct definition?, the validity evidence based on survey content was obtained through two stages; survey development and a pilot study. The results are summarized below.

The construct of effective teaching was delineated based on an extensive review of theory, research, and practice to develop the student rating survey. Thirty relevant studies were used to identify the research-based dimensions that were found to impact teaching quality and student learning. When combining these dimensions with three conceptual-based criteria of

AUN-QA, 7 dimensions were extracted. An initial version of the survey consisted of 24 items assessing 7 dimensions, as well as one global item measuring overall teaching effectiveness.

Expert opinions were sought to examine how adequate the dimensions are in covering teaching quality, how well the items are in assessing the dimensions, and how appropriate the selected response scale is. Furthermore, all items were checked to ensure the wording was clear, and there were no "double-barreled" items. This could also provide validity evidence based on response processes. The findings from the quantitative review of a panel of experts were as follows:

1. The survey items were relevant in assessing teaching quality, except for item 8 (the instructor adds information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline). It was removed from the survey,

2. The survey items assess the corresponding dimension, except for item 8,

3. The survey items were comprehensive in representing teaching quality,

4. The response scale was appropriate to use for the survey items.

The findings from the qualitative review showed that the number of dimensions needed to be refined. Four dimensions were adopted instead of seven dimensions because there was an overlap among proposed dimensions. In addition, two items were considered insufficient to assess a dimension. Consequently, the survey items were regrouped. Finally, item wording was changed to make items as simple as possible. A revised version of the student survey included four theoretical dimensions: (1) organization and structure, (2) assessment and feedback, (3) personal interactions, and (4) academic rigor. There was a total of 22 survey items and one global item. Results regarding the internal structure of the student survey are presented in the next section.

4.3 INTERNAL STRUCTURE OF THE SURVEY

4.3.1 Exploratory factor analysis (EFA)

An exploratory factor analysis (EFA) was used to examine the underlying construct in the dataset to offer insight into the internal structure. Even though the student rating survey was developed based on a confirmatory manner, it was useful to analyze data without prior assumptions.

EFA was performed on the revised 22 items of the student survey of instruction using Mplus (Muthén & Muthén, 1998-2012). The items were measured on a 5-point Likert scale, ranging from 1 (hardly at all) to 5 (to a very high degree). All items were positively worded items. The data was not multivariate normal, Mardia's skewness = 45.61, χ^2 = 17007.81, p < .001; Mardia's kurtosis = 1007.78, χ^2 = 348.91, p < .001. Because of the skewness in response distributions, robust weighted least squares extraction method (WLSMV) was used.

According to the Kaiser criteria, only one extracted factor had eigenvalues higher than 1 (Table 16). The result from the scree plot (Figure 3) also indicated that there was a 1 factor solution.

Table 16. Eigenvalues from the exploratory factor analysis of 22 items of the student rating survey (n = 2,234)

Factors	Eigenvalues	Ratio of Subsequent Eigenvalues
1	18.532	
2	0.368	0.020
3	0.328	0.891

Factors	Eigenvalues	Ratio of Subsequent Eigenvalues
4	0.279	0.851
5	0.224	0.803
6	0.199	0.888
7	0.176	0.884
8	0.170	0.966
9	0.164	0.965
10	0.162	0.988
11	0.157	0.969
12	0.149	0.949
13	0.140	0.940
14	0.129	0.921
15	0.126	0.977
16	0.117	0.929
17	0.114	0.974
18	0.108	0.947
19	0.098	0.907
20	0.095	0.969
21	0.087	0.916
22	0.077	0.885

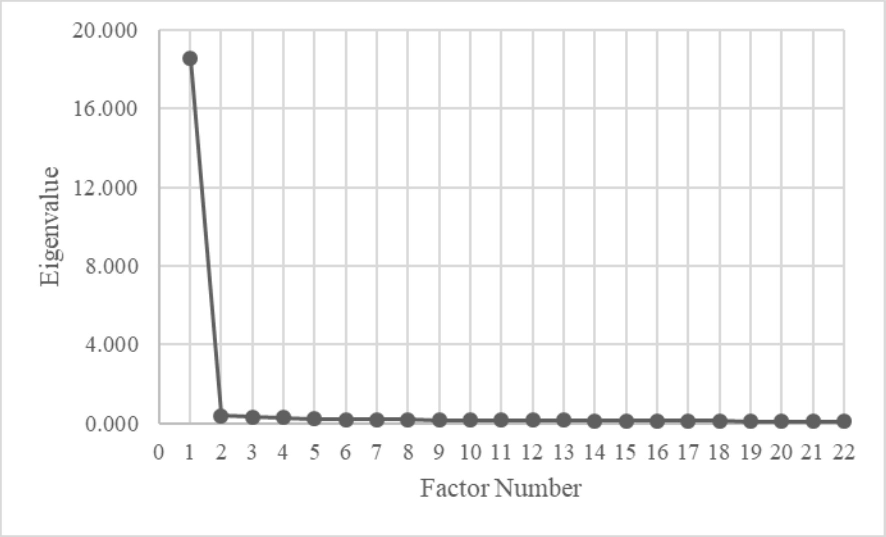


Figure 3. Scree plot for the exploratory factor analysis of 22 items of the student rating survey (n = 2,234)

There was a significant difference between observed and model correlation matrices, χ^2 (209, N = 2,234) = 2879.500, $p < .001$. However, since the model χ^2 is sensitive to sample size, four fit indices were examined to test model fit. Root Mean Square Error of Approximation (RMSEA) has cutoff values of less than .08 and .06 to reflect a reasonable and close fit to the data, respectively. Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) range from 0 to 1, and values greater than .90 and .95 support acceptable and excellent model fit, respectively. Lastly, standardized root mean square residual (SRMR) values of less than .08 reflect a close fit (Hu & Bentler, 1999; Mazor et al., 1992). CFI, TLI, and SRMR indicate a close/excellent fit, while RMSEA indicates a reasonable fit. Fit statistics for the 1-factor solution are provided below.

Table 17. Fit statistics for the initial 1-factor solution of the 22-item student rating survey

	One factor solution
χ^2 test of model fit	2879.500; *p* < .001
RMSEA	.076
CFI	.993
TLI	.992
SRMR	.015

RMSEA = root mean square error of approximation, CFI = comparative fit index, TLI = Tuker Lewis index, SRMR = standardized root mean square residual.

The first factor, which explained 84.24% of the variance of the item scores reflects one main underlying factor that represents general teaching competency. All 22 items loaded saliently onto the factor and salient loadings ranged from .892 to .927 (see Table 18).

Table 18. Factor loadings for 1-factor solution for 22 items of the student rating survey

Number	Item Stem	1-Factor Solution
Domain 1: Organization and Structure		
1	The instructor is well prepared for class.	.903
2	At the beginning of the course, the instructor specifies in detail course materials and grading procedures.	.905
3	The instructor presents the course content in an organized manner.	.911
4	There is close agreement between the announced	925

Number	Item Stem	1-Factor Solution
	objectives of the course and what is actually taught.	
5	The instructor explains the subject matter clearly.	.926
6	The instructor increases or improves my understanding of the subject matter.	.917
Domain 2: Personal Interactions		
7	The instructor creates a classroom atmosphere where I feel comfortable to express my ideas and ask questions.	.907
8	The instructor is interested in helping students learn.	.926
9	The instructor is friendly and establishes good rapport with students.	.914
10	The instructor keeps me engaged and helps me actively participate in productive learning.	.917
11	The instructor is reasonably available for help.	.892
Domain 3: Assessment and Feedback		
12	The instructor provides useful feedback that helps me understand my strengths and weaknesses.	.918
13	Graded assignments, tests, papers, homework, etc., are returned promptly.	.903
14	Grading in the course is fair and consistent.	.918
15	The exams reflect material emphasized in the course.	.927
16	Assigned work is appropriate to course level and credits.	.923
Domain 4: Academic Rigor		

Number	Item Stem	1-Factor Solution
17	The course is intellectually stimulating.	.925
18	The instructor uses a variety of instructional media/technology in class to enhance learning when applicable.	.901
19	The instructor attempts to stimulate creativity.	.927
20	The instructor encourages me to work and think independently.	.919
21	The instructor encourages me to apply the knowledge learned in this class to my work or other non-class related activities.	.924
22	The course encourages me to read further in the area.	.904

4.3.2 Justification for score reporting

It appeared that a one factor solution was justified. Based on the results, the evidence for reporting a total score was supported because of the large eigenvalue for a single factor (18.532). That is, the first eigenvalue was 50.4 times larger than the second. Also, the fit statistics indicated a good fit. Consequently, further analyses to investigate the multidimensionality of the survey responses were not conducted.

To argue that a total score is another measure of the global construct, the relationship between the summed scores and the scores from the global item was examined. A Spearman rank-order correlation was computed because the response distributions were skewed. There was a significantly positive strong relationship between the summed score and the global item, $r_s(2232) = .662$, $p < .001$.

Additional further validity evidence for a total score was examined in subsequent sections. A full discussion of benefits and shortcomings of reporting subscale scores and a total score is provided in Chapter 5.

4.3.3 Reliability

Although one general underlying factor representing teaching effectiveness was obtained, the Cronbach's alpha was calculated for each subscale as well as the total score (Table 19). It may be the case that some individual instructors in the sample differ on the domains to some extent. If some instructors do have differences, they may want to consider them when reflecting on their teaching. However, caution is needed when interpreting subscale scores because in the entire sample the survey was unidimensional.

Table 19. Cronbach's alpha coefficients for each subscale and the total score of the survey

	Number of items	Cronbach's alpha
Domain 1: Organization and Structure	6	.961
Domain 2: Personal Interactions	5	.950
Domain 3: Assessment and Feedback	5	.955
Domain 4: Academic Rigor	6	.961
Total score	22	.988

Generally, alpha coefficient levels above .7 show an adequate level of internal consistency. The Cronbach’s alpha in Table 18 reflected that each subscale and total score had high internal consistency.

4.3.4 Item-total score correlation

Based on the empirical testing, four dimensions initially proposed for the student rating survey were not supported. Thus, the item-to-total score correlations were only examined and the item-to-dimension correlations were disregarded. A polyserial correlation was computed between the scores of each survey item (polytomous ordinal variable) and the total scores (continuous variable) (see Table 20).

Table 20. Item-to-total score polyserial correlation for the student rating survey

Item	Polyserial correlation	Item	Polyserial correlation
1	.395	12	.402
2	.410	13	.386
3	.416	14	.396
4	.417	15	.414
5	.429	16	.405
6	.429	17	.405
7	.382	18	.376
8	.385	19	.401
9	.363	20	.377
10	.404	21	.398
11	.381	22	.418

The results showed that all correlations were positive. With a correlation of more than .3, it indicates very good discrimination.

4.4 RELATIONSHIP WITH OTHER VARIABLES

Validity evidence based on relations to other variables was obtained from the pilot study and the full implementation study. The relationship between student ratings and another measure of instructional quality (i.e., result from the previous student rating form used at Bangkok University) was obtained first in the pilot study. A Spearman rank-order correlation coefficient was computed to assess the relationship between the proposed survey and the Bangkok University current survey used to measure teaching effectiveness. There was a positive correlation between the two measures, r_s = .725, n = 109, p < .001. The correlation indicated a moderate positive relationship.

In the full implementation study, the relationships between student ratings and a measure of student achievement (i.e., overall GPA) was obtained for the entire sample. SPSS 23.0 software (IBM Corp., 2015) was used to evaluate this correlation.

In this study, 2,234 students completed the student rating survey and scores ranged from 22 to 110 (*mean* = 89.54, *SD* = 19.428). The distribution of scores was negatively skewed, with a value of -1.541. Student GPAX ranged from 0.66 to 4.00 (*mean* = 3.03, *SD* = .680) and the distribution of student GPAX was negatively skewed, with a value of -.648 (see Figure 4). Therefore, a Spearman rank-order correlation was computed to assess the relationship between the scores from student rating survey and student GPAX as a measure of student achievement. There was a significant small positive relationship between student ratings and student GPAX, r_s(2232) = .122, p < .001.

The result of the Spearman correlation showed a small positive relationship between student rating and student achievement. When compared to the correlation coefficients Cohen described the present coefficient is within the lower end of the range. The sample size of this study was moderately large and as mentioned earlier, the moderate-to-high correlations reported by Cohen appeared to derive from small sample size studies from particular courses.

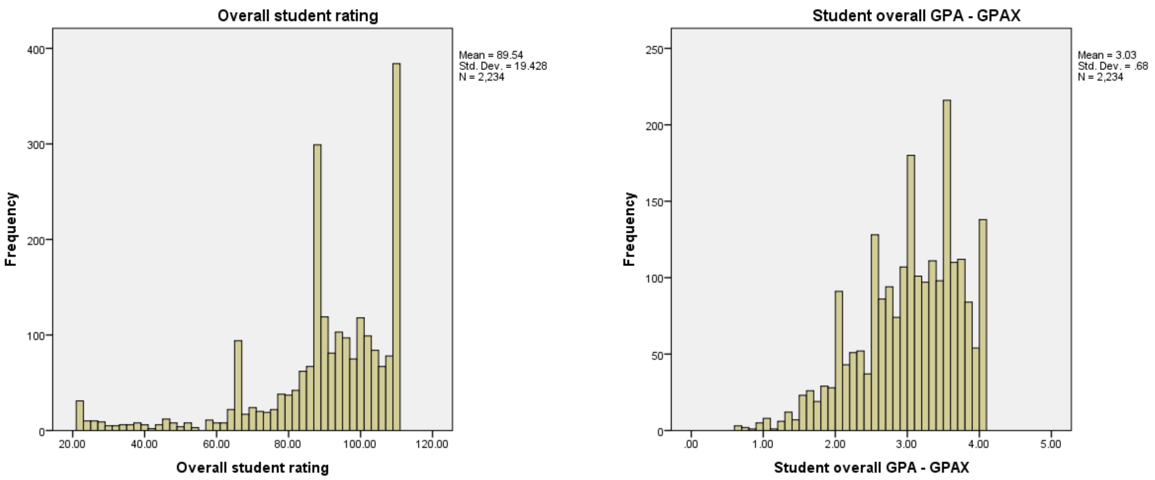


Figure 4. Frequency distribution of overall student rating and student overall GPA (n = 2,234)

4.5 DIFFERENTIAL ITEM FUNCTIONING

The DIFAS program (Penfield, 2005) was used to detect whether each item in the student rating survey exhibits uniform or non-uniform DIF between male and female students. For ordinal items, the generalized Mantel-Haenzel procedure is superior if a pattern of DIF is non-uniform. Whereas, the Mantel procedure should be used if a pattern of DIF is uniform. In this study, male was the reference group and was coded as “1”, while female was the focal group and was coded as “2”. The total score was used as a matching variable and was divided into ten equal intervals to match male and female students. The GMHDIF program (Fidalgo, 2011) was used to calculate the GMH chi square statistic along with a test of significance. Finally, the DIFAS program (Penfield, 2005) was used to calculate an estimator of the common odds ratio $(\hat{\alpha}_{LA})$ for the purpose of DIF classification.

4.5.1 Descriptive statistics for male and female students

There were 22 5-point Likert survey items assessed for DIF. The total scores ranged from 22 to 110. Overall, female students (*N* = 1,356, *Mean* = 90.59, *SD* = 18.578) rated instructors/courses slightly higher than male students (*N* = 878, *Mean* = 87.93, *SD* = 20.578). Table 21 displays descriptive statistics for the total rating scores for male and female students.

Table 21. Descriptive statistics of the total rating scores for male and female students

		N	*Mean*	*SD*	*Skewness*	*Kurtosis*
Gender	Males	878	87.928	20.578	-1.414	1.898
	Females	1,356	90.589	18.578	-1.625	3.176

4.5.2 DIF analyses

First, the ICCs from the rating scale model were used to examine the pattern of DIF by the DIFAS program (Penfield, 2005). The type of DIF present was identified by examining the differences between the difficulty indices for each subgroup. These differences are labeled as “the conditional differences”. If there is the same positive or negative direction across all the intervals of item performance, the DIF pattern is uniform. Conversely, the DIF pattern is non-uniform if the direction changes across the matching variable continuum. The matching variable (i.e., the total score) was classified into ten equal intervals. Table 22 presents the conditional differences in the mean survey item scores between male and female students at ten equal intervals across the matching variable continuum. Positive values indicate DIF in favor of the reference group, and negative values indicate DIF in favor of the focal group. Regarding its magnitude, the higher values of the conditional differences indicate that male and female students perform more differently in answering the same survey item. The smaller values indicate that male and female students perform less differently in answering the same survey item. The expectation value of the conditional differences in case of no DIF is zero. From Table 21, there are a mix of positive and negative values of the conditional differences. That means the probability of endorsing an item category for a specific item between males and females is not the same across ten levels of the total score. Hence, the survey items exhibit non-uniform DIF. The output from the DIFAS program is provided in Appendix C.

Table 22. Conditional differences in mean item score between the reference and focal groups at ten intervals

Lower	22.0	30.8	39.6	48.4	57.2	66.0	74.8	83.6	92.4	101.2
Upper	30.8	39.6	48.4	57.2	66.0	74.8	83.6	92.4	101.2	110.1
Item 1	0.02	-0.41	0.13	0.15	-0.16	-0.57	-0.01	0.03	-0.03	0.03
Item 2	0	-0.13	-0.07	0.45	-0.05	0.19	0.02	0.02	-0.08	0.03
Item 3	-0.04	-0.65	0.13	0.34	0.10	0.32	0.17	0.03	-0.03	0.01
Item 4	0.02	0.10	-0.13	0.34	-0.10	-0.01	-0.04	0.05	-0.02	0.01
Item 5	-0.10	-0.17	0	0.17	-0.03	0.10	-0.02	0.08	-0.02	0
Item 6	0.01	-0.25	0.13	-0.19	0	0.01	0.10	0.07	0.05	0.04
Item 7	-0.05	0.12	-0.20	0.43	-0.06	-0.20	-0.15	-0.07	-0.11	0
Item 8	-0.02	-0.05	-0.07	0.05	-0.11	0.09	-0.27	-0.03	0.05	0.04
Item 9	0.04	-0.02	0.27	-0.72	-0.03	0.29	0.03	-0.12	-0.07	0
Item 10	0.01	0.10	0.33	-0.43	-0.08	0.16	0.02	0	0.01	-0.01

Lower	22.0	30.8	39.6	48.4	57.2	66.0	74.8	83.6	92.4	101.2
Upper	30.8	39.6	48.4	57.2	66.0	74.8	83.6	92.4	101.2	110.1
Item 11	-0.05	0.15	0.33	0.15	0.10	0.23	0.05	-0.03	0.09	0.05
Item 12	-0.04	0.16	0.13	-0.13	0.11	0.06	-0.05	0.02	-0.05	0.02
Item 13	-0.01	0.24	-0.13	-0.22	0.04	0.05	0.01	0.03	0.05	0.04
Item 14	-0.04	-0.16	-0.07	0.07	0.08	0.03	0.07	0.01	0.05	0.03
Item 15	-0.05	-0.09	0.07	0.05	-0.06	-0.05	-0.13	0.01	0.10	0.03
Item 16	0.05	-0.03	-0.33	-0.30	0.10	0.13	-0.04	-0.02	0	0.01
Item 17	-0.10	-0.19	0.20	-0.05	-0.08	-0.02	-0.05	0.05	-0.01	0.05
Item 18	0.03	0.12	0.07	-0.14	0.06	-0.08	-0.14	-0.03	-0.06	0.01
Item 19	-0.06	0.06	-0.07	0.43	0.07	-0.13	-0.12	-0.06	-0.09	0
Item 20	0.08	0.15	-0.33	-0.35	0.07	0.16	-0.12	-0.04	-0.01	0
Item 21	0.08	-0.22	-0.20	0.04	0.05	-0.12	-0.10	-0.08	0.04	-0.01
Item 22	0.04	0.11	-0.07	-0.54	-0.08	-0.26	-0.08	-0.02	0.01	0.05

From Table 22, the results suggest that 22 survey items in the student rating survey may exhibit non-uniform DIF. Therefore, the generalized Mantel-Haenzel procedure was used to analyze DIF and to answer the third research question (is there gender differential item functioning in student ratings?).

The GMHDIF program (Fidalgo, 2011) was used to calculate the generalized Mantel-Haenzel statistics (χ^2_{GMH}) to test the null hypothesis of no association between subgroups (males vs. females) and item categories. A two-stage GMH process was performed. If there are any items found to exhibit DIF in the first stage, they are removed for calculating the matching criteria in the second stage. An item is free of DIF if a *p*-value is more than .05. In the case of H_0 rejection, DIF is found on the studied item. As is always the case, statistical significance does not imply practical importance. DIF that is a small fraction of a score point in magnitude may not be worrisome even if it is statistically significant. In judging importance of DIF based on both statistical significance and the size of the DIF index, an estimator of the Liu-Agresti cumulative common log-odds ratio $(\hat{\alpha}_{LA})$ of each item was calculated using the DIFAS program. Table 23 displays the results of the GMH chi-square statistic and the Liu-Agresti estimator of an odd ratio in the first stage.

Table 23. DIF result of the first stage GMH analysis between male and female students

Item	χ^2_{GMH} or $Q_{GMH(1)}$	*p*-value	$\hat{\alpha}_{LA}$
1	5.040	.283	-.111
2	2.621	.623	-.051
3	7.780	.100	.161
4	2.396	.663	.039
5	2.908	.573	.067
6	9.327	.053	.256
7	9.267	.055	-.330
8	3.084	.544	-.054
9	5.080	.279	-.239
10	1.106	.893	-.018
11	5.688	.223	.245

Item	χ^2_{GMH} or $Q_{GMH(1)}$	*p*-value	$\hat{\alpha}_{LA}$
12	1.414	.842	.030
13	5.283	.259	.169
14	7.255	.123	.173
15	9.107	.058	.130
16	1.768	.778	.052
17	7.989	.092	.103
18	8.560	.073	-.115
19	5.526	.238	-.219
20	1.558	.816	-.069
21	6.512	.164	-.176
22	10.034 *	.040	-.075

* indicate *p*<.05

In the first stage, only the GMH chi-square statistic of item 22 was significant, indicating an association between subgroups (males vs. females) and item categories. According to the ETS classification scheme, item 22 was flagged as an "*A*" item (exhibiting negligible DIF) as $|\hat{\alpha}_{LA}| <$ 0.43.

A second stage of DIF was carried out using a pure matching variable. Item 22 which had been detected with DIF in the first stage was eliminated from the total score. All items were analyzed again and the results was provided below (Table 24).

Table 24. DIF result of the second stage GMH analysis after excluded DIF item between male and female students

Item	χ^2_{GMH} or $Q_{GMH(1)}$	*p*-value	$\hat{\alpha}_{LA}$
1	4.783	.310	-.111
2	2.963	.564	-.051
3	6.738	.150	.161
4	1.565	.815	.039
5	3.664	.453	.067
6	7.427	.115	.256
7	9.934*	.042	-.330
8	3.324	.505	-.054
9	6.365	.173	-.239
10	0.767	.943	-.018
11	6.639	.156	.245
12	1.403	.844	.030
13	5.425	.246	.169
14	7.812	.099	.173
15	7.744	.101	.130
16	1.897	.755	.052
17	8.084	.088	.103
18	7.733	.102	-.115
19	6.314	.177	-.219
20	3.353	.500	-.069

Item	χ^2_{GMH} or $Q_{GMH(1)}$	*p*-value	$\hat{\alpha}_{LA}$
21	3.801	.434	-.176
22	10.034 *	.040	-.075

* indicate *p*<.05

Note that the chi-square of item 22 remained the same, while other chi-squares changed. This is because the matching score was based on DIF-free items plus the studied item. In Table 22, all items contributed to the total score but after the first stage DIF was analyzed, item 22 was removed from calculating the total score in the second stage. However, calculating chi-square for item 22 in the second stage still needed to include the studied item. Hence, there was no change in the chi-square of item 22.

From Table 24, items 7 and 22 were identified as exhibiting DIF based on significant tests of GMH chi-square statistic. Item 7 had a significant χ^2_{GMH} value of 9.934 (*p* = .042). Item 22 had a significant χ^2_{GMH} value of 10.034 (*p* = .040). The $\hat{\alpha}_{LA}$ for these two items were -.330 and -.075, respectively. However, they both were classified as "*A*" items, showing negligible DIF as $|\hat{\alpha}_{LA}| < 0.43$, favoring females (the focal group) as $\hat{\alpha}_{LA}$ were negative values. Item 7 pertains to the extent the instructor creates a classroom atmosphere where students feel comfortable to express their ideas and ask questions. Item 22 pertains to the extent the course encourages students to read further in the area.

4.5.3 Summary of DIF analyses

Overall, two out of twenty-two items on the student rating survey were flagged as "*A*", showing negligible DIF between male and female students based on the results from the DIF analyses. Both items favored females, indicating that these items were slightly easier to get high rating by females.

According to ETS classification criteria, an item with negligible DIF or an item classified as *A* can be considered as no DIF. Only an item classified as *B* or *C* needs to be reviewed for potential bias and revised or removed. Hence, the meaning of survey categories/scales was shared across male and female students from this study. That is, there is some evidence that they understood and interpreted the survey scales in the same way and no survey item was removed after DIF analyses.

5.0 DISCUSSION

The main purpose of this study was to develop and validate a student rating survey that accurately and reliably measures teaching effectiveness. Validity evidence was gathered from three phases of research; survey development, a pilot study, and a full implementation study. Evidence for each research question are discussed in the following sections.

5.1 DISCUSSION OF RESEARCH QUESTIONS

5.1.1 Research question 1: To what extent does the content evidence support the construct definition?

From the literature, student rating surveys vary greatly in terms of the content and the number of dimensions because there is no single criterion to represent good teaching. The literature review suggested that both theory and empirical testing are needed to develop useful student rating surveys (Abrami, d'Apollonia, & Cohen, 1990; Apodaca & Grad, 2005; Marsh & Hocevar, 1991; Richardson, 2005; Spooren et al., 2013).

In this study, validity evidence based on survey content was acquired through the survey development and the pilot study. The survey development involved many phases that were designed to gather the data needed for validity evidence. Finally, the pilot study was conducted to provide additional validity evidence.

First, the purposes and uses of student ratings and the construct of teaching quality were delineated. An extensive review of research examining dimensionality of student ratings of instruction in higher education was conducted to identify research-based dimensions that were consistently found to impact teaching quality and student learning. These dimensions were combined with three conceptual-based criteria of AUN-QA that were used to facilitate the assessment of teaching quality in ASEAN universities. The survey items were then written/selected in a way that was consistent with their corresponding dimensions. The appropriate response scales were developed based on logical and technical requirements to make the survey items as objective as possible. Specifically, the response scales connect the inference into observable performance and provide information for evaluating teacher performance. Hence, it is important to apply logical and technical requirements when developing response scales.

All dimensions, survey items, and response scales were reviewed using content validity criteria (Armstrong et al., 2005; Thrush et al., 2007; Wynd et al., 2003). It involved experts' judgments which provided a logical analysis of the relationship between the content of student ratings of instruction and the construct being measured. The results showed that the dimensions and survey items were perceived adequate in covering teaching quality, the survey items were perceived to properly assess the associated dimensions, and the response scales were perceived suitable with what they were intended to measure. Only one item was removed due to construct under-representation, and the number of dimensions was refined based on the qualitative review. Four dimensions were adopted instead of seven dimensions because there was an overlap among proposed dimensions. The four dimensions were “Organization and Structure”, “Assessment and Feedback”, “Personal Interactions”, and “Academic Rigor”.

Next, the pilot study was conducted using the revised student survey. The results provided some evidence in support of the overall construct definition but not the dimensions. All survey items highly correlated with each other. This most likely suggests that the survey was measuring the same construct of teaching effectiveness for this sample of students. With regard to the response scale, students applied the full range of the response scale to the performances in the classroom for 13 out of 22 survey items. The response category labeled as “hardy at all” was not chosen by students for 9 items. However, due to a very small sample size of the pilot study, the response scale was maintained for the full implementation study with the larger sample size. It should be noted the distribution of the ratings were negatively skewed.

In summary, the development of this student survey was based on logical and empirical analyses and was guided by theory, empirical research, and practice from the University. Although the expert review indicated that the items were measuring the intended dimensions, the empirical results suggest that the construct of teaching effectiveness as defined in this survey may be unidimensional.

5.1.2 Research question 2: To what extent do the relationships among survey items and survey components correspond to the construct dimension?

With respect to validity evidence based on internal structure, it is important to examine the relationships among survey items and theoretical survey components. These relationships must represent the construct of teaching quality on which the proposed survey score interpretations and uses are based (AERA, APA, NCME, 2014). In this study, student ratings of instruction are used for both formative and summative purposes, that is to provide feedback to teachers for instructional improvement, to provide input for administrative decision making, and to provide evidence for demonstrating the performance of an institution. For instructional improvement, the scores from several dimensions are needed to provide feedback to faculty in terms of specific areas in need of improvement. For administrative decision making and demonstrating the performance of an institution, a summed score from a survey is needed as a measure of overall teaching effectiveness. The theoretical framework for teaching effectiveness can be considered either unidimensional or multidimensional. However, the literature review showed that the multidimensionality of teaching is widely accepted. Hence, the development of this student survey was based on a multidimensional framework.

To examine the internal structure of the student survey, factor analysis and relationships between survey items and dimensions were studied. Validity evidence based on internal structure was discussed in relation to the pilot study and the full implementation study.

Factor analysis

Exploratory factor analysis was performed on an entire sample of the full implementation study to offer insight into the internal structure of the survey. The data were not multivariate normal therefore the robust weighted least squares extraction method was used.

A one-factor solution was extracted. All 22 items from the four theoretical dimensions loaded saliently together onto one factor. A recent study of Spooren et al. (2013) revealed that many dimensions in student rating surveys seem to be affected by a global (unidimensional) construct. This factor most likely is a general factor that represents general teaching competency. Although the hypothesized multidimensional factor structure was not supported by the data, it may be the case that the results across the four dimensions vary for some instructors.

Based on the results from the factor analysis, there was one dominant eigenvalue, supporting a total score instead of subscale scores. A total score could provide an overall teaching competency to support the interpretation entailed by summative proposed uses of the survey. However, the use of subscale scores is important when improving or changing specific classroom practices. Even though the results did not support multidimensionality, the dimensions can still be used by individual instructors to evaluate their own teaching.

Relationships between survey items and total/dimension scores

The polyserial correlation between survey items and total scores ranged from .853 to .959 in the pilot study. With a correlation of more than .3, it indicates very good discrimination and none of the survey items were excluded from the analyses. However, the more each item correlates with the survey score as a whole, the higher all items correlate with each other. This most likely suggests that the survey is unidimensional for this pilot study sample.

In terms of polyserial correlations between survey items and theoretical dimensions, the item-dimension correlation of each item showed the highest value with its theoretical dimension. However, the correlations with other dimensions were still very high. Thus, the proposed multidimensional model of university teaching to the pilot study sample may not hold.

For the full implementation study, only the item-to-total score correlations were examined because the empirical testing showed that the four dimensions initially proposed for the student rating survey appeared to be affected by a global (unidimensional) construct. The results showed that all correlations were positive ranging from .363 to .429. With a correlation of more than .3, it indicates very good discrimination.

In the literature, validity studies of internal structure focused on the relationships among survey items and dimensions by presenting item-total score correlations and item-dimension (subscale) score correlations. If the survey items are written in a way that is consistent with their corresponding domain, the correlations of the survey items that belong to the same underlying domain yield higher correlations than those belonging to other dimensions. (Bell et al., 2012; Spooren et al., 2013). This conclusion is consistent with the results from the pilot study. The item-dimension correlation of each item showed the highest value with its theoretical dimension. Likewise, if the survey items are affected by a global construct, the item-total score correlations yield high relationships with the survey score as a whole.

Reliability

Reliability was examined during the pilot study and full implementation study. Cronbach's α coefficients were computed to support internal consistency of the total survey score and all subscales scores. For the pilot study, the results showed strong support. Cronbach's α coefficients of each subscale score were extremely high (>.9). For the full implementation study, Cronbach's α coefficients of each subscale score and total score were also extremely high (>.9).

The literature has shown that when outliers are taken into account, the estimates of Cronbach's coefficient alpha are seriously inflated (Y. Liu & Zumbo, 2007). However, ordinal item response data is rarely affected by outliers. Zumbo et al. (2007) found that the estimates of coefficient alpha computed from ordinal response data are downward biased when compared with one computed from an interval/ratio scale, especially when skewness in response distribution is involved. Highly skewed items have been found to decrease estimated reliability when there are fewer response categories. All survey items within the student rating survey were negatively skewed, with more cases in the higher end of the response scale. Another study of

Zumbo et al. (2007) concluded that the bias on reliability estimates decreased when the number of scale points increased, but increased when negatively skewed items were used. In this case, high α coefficients computed from the samples of the pilot study and full implementation study are more likely appropriate, or even underestimated. Therefore, strong support for reliability was obtained.

5.1.3 Research question 3: Is there gender differential item functioning in student ratings?

Another validity concern that can affect survey score interpretation is how students interpret the scale. For ordinal scales, there is a chance that students share a different understanding toward the same item. This problem arises when data is analyzed and interpreted, because the meaning of coded values is not shared across subgroups. This concern can be reflected in Differential Item Functioning (DIF) (Granberg-Rademacker, 2010).

Gender DIF has been detected in many instruments/tests. Yet, this issue in student ratings of instruction has not been fully resolved. Earlier research provided conflicting results regarding the relationship between the gender of the students and student ratings of instruction. A number of research studies (Beran & Violato, 2005; McPherson & Jewell, 2007; McPherson et al., 2009) reported no differences between faculty ratings made by male and female students. This may indicate no DIF items. Other studies reported differences in ratings given by male and female students. This may indicate the occurrence of DIF.

To answer the question of gender differential item functioning in student ratings, DIF was performed using the two-stage GMH process. As the results demonstrated, two out of twenty-two items on the student rating survey were flagged as *A*, showing negligible DIF between male and female students. Both items favored females, indicating that at least one of the higher-end category responses was more easily endorsed by females given that they had the similar levels on the intended survey attribute. In other words, these items were slightly easier to get high ratings by females. Centra and Gaubatz (2000) and Kohn and Hatfield (2006) examined the impact of gender on student ratings of instruction and found that female students generally gave their teachers higher ratings than male students. However, they were examining overall impact differences not DIF.

According to ETS classification criteria, an item with negligible DIF or an item classified as *A* can be considered as no DIF. Hence, both items still remain in the survey. Overall, the results indicate that there is no occurrence of DIF in this student survey. The validity evidence of survey score interpretations was supported since the meaning of survey categories/scales was shared across male and female students. That is, there is some evidence that they understood and interpreted the survey scales in the same way.

However, the issue regarding sample size and sample ratio need to be addressed. In this study, 878 male and 1,356 female students were used to detect DIF. Ryan (2008) studied how sample size and sample ratio influenced power of GMH. He found that the highest power for GMH procedure was discovered when a sample size was 1,000 and with a 50:50 sample size ratio between subgroups. In this case, a random sample of 1,000 male students and 1,000 female students is needed to obtain the highest power. Unfortunately, this case could not be done in this study because there was not enough of male students. To assure the highest power, a further investigation with larger sample sizes is needed.

5.1.4 Research question 4: Are there relationships between student ratings and a similar measure of teaching quality and student achievement?

External validity evidence was obtained to answer this research question. This source of validity evidence can support the extrapolation inference to make claims about whether the scores from student ratings of instruction are related to the target domain (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Bell et al., 2012; adapting from Kane, 2006). A similar measure of teaching quality was expected to relate with student ratings of instruction. In addition, student achievement was expected to be predicted by the student ratings based on a series of articles by Cohen (1980, 1981, 1982).

First, convergent validity evidence was reported by examining the relationship between student rating scores and scores from a survey currently used at Bangkok University. Correlation analysis was used to detect this relationship during the pilot study. Both surveys measured similar intended construct of teaching quality. Hence, a positive relationship was expected to support the proposed interpretation of the survey scores. Overall, student rating scores showed a significant relationship of .725 with scores from a survey currently used at Bangkok University. There was a moderately strong positive relationship between the two measures. This magnitude of relationship was not surprising. It is important to note that only 109 students participated in the pilot study. These students volunteered to be in the pilot study. They received the link of the online student rating survey through their university emails during their final examination period. Whereas, the current survey used at Bangkok University was open during the last three weeks of teaching and two weeks of final examination period. Students needed to log into the Bangkok University Online Assessment System with their university ID number and password to complete the survey. The differences in administration methods (email vs. Bangkok University Online Assessment System) and administration timeframes (final examination period vs. last three weeks of teaching and two weeks of final examination period) might have affected students' ratings.

Second, the relationship between student ratings and student achievement was examined during the full implementation study for providing further external validity evidence. This relationship is the common method in many published studies to collect external validity evidence for student ratings of instruction. Student achievement is considered as the most commonly used criterion variable in many studies (Onwuegbuzie et al., 2009; Ory & Ryan, 2001). A series of articles by Cohen (1980, 1981, 1982) synthesized validity studies and concluded that the relationship between overall student ratings and student GPA as a measure of students' achievement should be positive. This positive effect can imply that high instructor/course ratings should predict high student achievement. However, student achievement for some of the studies was measured by a course GPA not an overall GPA. The use of a course GPA versus an overall GPA may have an impact on the correlations between student achievement and student ratings.

For the full implementation sample, the correlation between student rating scores and students' overall GPA as a measure of student achievement was significantly positive but weak. This .122 positive relationship was as expected but the magnitude was small. As previously mentioned, the overall relationship between student ratings and student achievement varied in the studies summarized by Cohen. Large and moderate correlations were obtained in the small sample sized studies, while no or weak correlations were collected from the large sample sized studies (Cohen, 1980, 1981, & 1982). The sample size of this full implementation study was fairly large with *n* = 2,234. It would be possible for a large group of high GPA students (and low GPA students) across various disciplines and courses to view the instructor/course differently.

A recent meta-analysis study of Uttl, White, and Gonzalez (2017) revealed no or minimal significant correlations between student ratings and learning. These findings suggest that institutions should not focus on student achievement solely when evaluating the instructors/courses. Instead, a student rating survey can be used to understand students' perceptions toward instructors/course.

Finally, the results based on relation to other variables showed strong positive relationship between the student survey and another currently used survey at Bangkok University which was used to evaluate teaching effectiveness for a decade. This could indicate that the student survey was also measuring a similar construct to teaching effectiveness.

The student survey captures students' perceptions toward instructors/course and can be used as a tool for formative purposes. Results from the student survey can be used by the instructors to improve or change classroom practices and eventually this may impact student performance.

5.2 LIMITATIONS AND FUTURE DIRECTIONS

There are limitations in this study which in turn support future research. The use of a small sample size in the pilot study is the first limitation to be addressed. Only 109 students were included in the pilot study. This small sample size issue actually occurred due to a low response rate, which is addressed more in details in the second limitation. Still, this is considered low for validation studies and some of the quantitative analyses. In the future, a larger sample should be randomly or purposely selected so that the sample size is more appropriate for some of the quantitative techniques. Considering the sample size of the full implementation study, a total of 2,234 students is not considered small. However, with a 50:50 sample size ratio between subgroups (i.e., males and females), there were not enough male students to assure the highest GMH's power to detect DIF.

The second limitation of this study encompasses low response rates. During phase IV of survey development, the items were subjected to a review conducted by an external panel of academic leaders, instructors, psychometricians, as well as students from Bangkok University, Thailand. There was a total of 81 expert panelists voluntarily participating in the review. The questionnaire with seven items was sent to them to complete. Only 55% of the total panelists completed item 2, while approximately 85% of the total panelists completed the other items. After inspecting the questionnaire, it was found that item 2, the problematic item, was located in the second column of the table in the questionnaire. While, item 1 was located in the first column. Without careful consideration, the panelists might only complete item 1 and miss item 2. Another concern regarding low response rate occurred during the pilot study. An overall response rate of 17.33% was obtained which may have affected the results of the pilot study.

The third limitation is related to skewed item responses. There were some very low frequencies on the lower-end of response categories. Typically, the distributions of student ratings of instruction are negatively skewed. That is, students tend to rate their instructors/courses high. This skewed distribution could limit the choice of the quantitative techniques for data analysis. It is hard to detect if this negative skewness in item response truly reflects students' perceptions toward their instructors/courses (which likely occurred in the majority of the research) or it reflects a ceiling effect in survey scores. There was evidence that some students were responding with the same category of “to a very high degree” across items. Responding to the survey in this way may accurately reflect the student’s perception of the instructor/teaching, but it may also indicate a response set. Future investigation on students’ positive perceptions toward their instructors/courses should be helpful to capture a more comprehensive picture of their responses.

The fourth limitation is related to the skewed distribution discussed earlier. To examine the pattern of DIF in this study, the total scores from the student survey were divided into ten equal intervals. Given that the distribution was skewed, another approach would have been to use unequal intervals to ensure a large enough sample size for the intervals at the lower end of the scale.

The next limitation is related to restriction in collecting five sources of validity evidence. Based on the 2014 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), five sources of validity evidence are recommended in supporting a validity argument of student ratings of instruction. Due to the time restriction, validity evidence based on survey consequences was not in the scope of this study. With so many institutions using student ratings in personnel decisions, the consequences of their use need to be addressed, including both positive and negative consequences. The examination of survey consequences pertaining to faculty personnel decisions and classroom practices should be examined.

Finally, although only a Thai version of the student survey was used by students in this study, the comparability of Thai and future English versions of the survey should be examined to support the validity of the results for both language versions.

5.3 CONCLUSION

Bangkok University, Thailand has adopted student surveys as a measure of teaching quality for several decades. In recent years, ASEAN University Network (AUN) has required ASEAN universities to change and adjust some criteria of teaching quality. Therefore, the development and validation of a student survey that met the requirements of the AUN framework was needed.

The new student survey was developed for three main purposes: (a) providing feedback to teachers for instructional improvement, (b) providing input for administrative decision making, that is faculty promotion and tenure decisions, and (c) providing evidence for demonstrating the performance of an institution as a part of internal quality assurance processes. It is important to ensure that the survey results are valid for its intended purposes. To provide meaningful and interpretable results, four sources of validity evidence were collected to support the content and construct domains of the survey. To this end, this study evaluated the extent to which the content evidence supported the construct definition of the survey (RQ1), the relationships among survey items and survey components corresponded to the construct dimension (RQ2), the survey exhibited gender differential item functioning (RQ3), and the relationships between student ratings and a similar measure of teaching quality and student achievement (RQ4).

Based on expert reviews, the dimensions and survey items were perceived adequate in covering teaching quality, the survey items were perceived to properly assess the associated dimensions, and the response scales were perceived suitable with what they were intended to measure. Exploratory factor analysis suggested that the construct of teaching effectiveness as defined in this survey may be unidimensional. Although the results did not support multidimensionality, it is only one form of assessing teaching effectiveness. The survey cannot capture the complexities of the teaching process but the dimensions can still be used by individual instructors to evaluate their own teaching. Cronbach's α coefficients were high and supported the internal consistency of the survey. In term of gender DIF, there was no occurrence of gender DIF in this student survey. Therefore, the validity evidence of survey score interpretations was supported since the meaning of survey categories/scales was shared across male and female students. Finally, the results based on relation to other variables showed a strong positive relationship between the student survey and another currently used survey at Bangkok University which was used to evaluate teaching effectiveness for a decade. This could indicate that the student survey was measuring a similar construct of teaching effectiveness.

In closing, the student survey demonstrated good psychometric quality, which was examined using classical analyses. All survey items indicated very good discrimination. Validity evidence was obtained to support the intended purposes and uses of the survey scores.

APPENDIX A

A COMPLETE LIST OF TEACHING DIMENSIONS EXTRACTED FROM THE LITERATURE REVIEW

Table A1. A summary of dimensions in literature review

Author	Number of dimensions	Dimensions
Barnes et al. (2008)	7	Preparedness Professionalism Evaluation Rapport Enthusiasm Delivery Excellence
Barth (2008)	5	Quality of instruction Course rigor Level of interest Grades Instructor helpfulness
Chatterjee, Ghosh, and Bandyopadhyay (2009)	8	Academic excellence and knowledge in the subject Personality, behavior and appearance Ability to teach and mode of presentation Use of technical gadgets and ability in providing updated information Availability in the department beyond class hour for discussion on academic matters

Author	Number of dimensions	Dimensions
		Regularity and punctuality
		Ability to communicate and impress student
		Ability to guide after passing out from the department
Coffey and Gibbs (2001)	11	Learning/value
Marsh (2007b)		Enthusiasm
Marsh et al. (2009)		Organization/clarity
		Group interaction
		Individual rapport
		Breadth of coverage
		Examinations/grading
		Assignments
		Workload/difficulty
		Overall course
		Overall instructor
Cohen (2005)	2	Course
		Teacher
Díaz, Swan, Ice, and Kupczynski (2010)	10	Design and organization
		Facilitation
		Direct instruction
		Affective expression

Author	Number of dimensions	Dimensions
		Open communication
		Group cohesion
		Triggering event
		Exploration
		Integration
		Resolution
Feldman (2007)	28	Teacher’s stimulation of interest in the course and its subject matter
		Teacher’s enthusiasm
		Teacher’s knowledge of subject matter
		Teacher’s intellectual expansiveness
		Teacher’s preparation, organization of the course
		Clarity and unpersuadableness
		Teacher’s elocutionary skills
		Teacher’s sensitivity to, and concern with, class level and progress
		Clarity of course objectives and requirements
		Nature and value of the course material
		Nature and usefulness of supplementary materials and teaching aids

Author	Number of dimensions	Dimensions
		Perceived outcome or impact of instruction
		Teacher’s fairness, impartiality of evaluation of students, quality of examinations
		Personality Characteristics of the teacher
		Nature quality, and frequency of feedback from the teacher to students
		Teacher’s encouragement of questions and discussion, and openness to opinions of others
		Intellectual challenge and encouragement of independent thought
		Teacher’s concern and respect for students, friendliness of the teacher
		Teacher’s availability and helpfulness
		Teacher motivates students to do their best, high standard of performance required
		Teacher’s encouragement of self-initiated learning
		Teacher’s productivity in research related activities
		Difficulty of the course (and workload) – description

Author	Number of dimensions	Dimensions
		Difficulty of the course (and workload) – evaluation Classroom management Pleasantness of classroom atmosphere Individualization of teaching Teacher pursued and/or met course objectives
Glazerman et al. (2011)	7	Understand subject matter Connect what is to be learned to students’ prior knowledge and experience Create effective scaffolds and supports for learning Use instructional strategies that help students draw connections, apply what they are learning, practice new skills, and monitor their own learning Assess student learning continuously and adapt teaching to student needs Provide clear standards, constant feedback, and opportunities for revising work Develop and effectively manage a collaborative classroom in which all students

Author	Number of dimensions	Dimensions
		have membership
Ginns, Prosser, and Barrie (2007)	5	Good teaching Clear goals and standards Appropriate assessment Appropriate workload Genetic skill
Gursoy and Umbreit (2005)	4	Organization Workload Instruction Learning
Keeley et al. (2006, 2010)	2	Caring and supportive Professional competency and communicational skills
Kelly, Ponton, and Rovai (2007)	20	Instructor's attitude Instructor's rapport Instructor's personality Instructor's knowledge Instructor's stimulation Instructor's ability Instructor's preparedness Instructor's helpfulness

Author	Number of dimensions	Dimensions
		Overall teacher
		Course’s organization
		Course’s content
		Course’s materials
		Course’s workload
		Course’s lecture
		Course’s discussion
		Course’s assignments
		Overall course
		Grading fairness
		Grading timeliness
		Quality/quantity feedback
Kember and Leung (2008)	9	Understanding fundamental content
		Relevance
		Challenging beliefs
		Active learning
		Teacher-student relationships
		Motivation
		Organization
		Flexibility
		Assessment

Author	Number of dimensions	Dimensions
Mortelmans and Spooren (2009)	12	Clarity of course objectives
Spooren (2010)		Value of subject matter
		Build-up of subject matter
		Presentation skills
		Harmony organization course-learning
		Course materials
		Course difficulty
		Help of the teacher during the learning process
		Authenticity of the examination
		Linking-up with foreknowledge
		Content validity of the examination
		Formative evaluation(s)
Renaud and Murray (2005)	8	Clarity
		Enthusiasm
		Interaction
		Organization
		Pacing
		Disclosure
		Speech
		Rapport
Roberts (2008)	21	Organization

Author	Number of dimensions	Dimensions
		Effective communication skills
		Interest in student progress
		Prompt feedback
		Freedom to think
		Course value
		Workload/difficulty
		Instructor overall
		Overall GPA
		Probable grade
		Clear objectives
		Content knowledge
		Interest in teaching/enthusiasm
		Fair evaluation
		Stimulate interest/motivating
		Active learning
		Evaluation of teaching/learning from students
		Use of supplementary materials
		Classroom management
		Individual rapport
		Class interaction
Safavi et al. (2012)	3	Learning enhancement

Author	Number of dimensions	Dimensions
		Interpersonal skills-overall
		Assessment and grading
Sedlmeier (2006)	9	Transparencies and other course materials
		Clarity of instruction
		Interestingness of subject matter
		Competence of instructor
		Relevance for exam
		Literature (readability, availability)
		Cooperation between instructor and students
		Atmosphere/rooms
		Instructor’s appearance
Shevlin et al. (2000)	2	Lecturer ability
		Module attributes
Simendinger et al. (2017)	5	Content delivery
		Behaviors, competences, and skills
		Structural and personal traits
		Cultural and socio-economic context
		Overall teaching effectiveness
Spooren, Mortelmans, and Thijssen (2012)	12	Clarity of course objectives
		Relevance of subject matter
		Build-up of subject matter

Author	Number of dimensions	Dimensions
		Presentation skills
		Harmony organization course-learning
		Course materials
		Course difficulty
		Help of the teacher during the learning process
		Authenticity of the examination
		Linking-up with foreknowledge
		Content validity of the examination
		Formative evaluation(s)
Toland and DeAyala (2005)	3	Instructor’s delivery of course information
		Teacher’s role in facilitating instructor/student interaction
		Teacher’s role in regulating student’s learning

APPENDIX B

A QUESTIONNAIRE USED TO CONDUCT A FIELD TRIAL IN PHASE IV

A development of the new survey for Bangkok University online assessment system

Direction: the items are ordered by the domains they represent. Please rate how well each item assesses the domain and how relevant each item assesses the teaching quality. What modifications are needed to improve the clarity and meaningfulness of the items?

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
D1: Planning and Preparation										
1. The instructor is well prepared for class. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
2. The instructor clearly explains the course objectives in the beginning of class. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
3. The instructor presents the course content in an organized manner. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
4. There is close agreement between the announced objectives of the course	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
and what is actually taught. Comment…………………………..										
D2: Classroom Management and Environment										
5. The instructor maintains a classroom atmosphere where I feel comfortable to express ideas and ask questions. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
6. The instructor uses appropriate teaching methods which helps my learning. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
7. The instructor uses a variety of instructional media/technology in class when applicable. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
8. The instructor adds the information related to ethics and morality to the teaching method, e.g., honesty, responsibility, discipline. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
9. The instructor raises challenging questions and problems. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
10. The instructor helps to keep me engaged and participated in productive learning. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
D3: Knowledge of Subject Matter										
11. The instructor is competent in his/her knowledge of subject. Comment...............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
12. The instructor increases or improves my understanding about subject matter. Comment..............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
D4: Clarity of Presentation										
13. The instructor explains the subject matter clearly. Comment..............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
14. The instructor refers to experiences or examples to clarify concepts. Comment..............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
D5: Availability and Helpfulness										
15. The instructor provides useful feedback that help me understand my strengths and weaknesses. Comment..............................	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
16. The instructor is reasonably accessible for help. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
17. The instructor provides feedback in a timely fashion. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
D6: Evaluation and Quality of Examination	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
18. The instructor evaluates my work fairly. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
19. The exams reflect material emphasized in the course. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
20. Assigned work is appropriate to course level and credits.	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
Comment………………………..										
D7: Student Outcome	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
21. The activities in class keep me interested and motivated. Comment………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
22. The instructor attempts to stimulate creativity. Comment………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
23. The instructor encourages me to work and think independently. Comment………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
24. The instructor encourages me to apply the knowledge created in this class to my work or other non-class related activities. Comment………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

	How well each item assesses the domain					How relevant each item assesses the teaching quality				
	Extremely well (1)	Very well (2)	Moderately well (3)	Slightly well (4)	Not well at all (5)	A great deal (1)	A lot (2)	A moderate amount (3)	A little (4)	None at all (5)
25. The instructor’s overall teaching effectiveness. Comment…………………………..	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

Overall comprehensiveness of the 25 survey items in representing teaching quality

○ Extremely (1)

○ Very (2)

○ Moderately (3)

○ Slightly (4)

○ Not at all (5)

Please provide additional survey item(s) within domains that should be added to improve the comprehensiveness of the survey

Domain 1: Planning and Preparation

Domain 2: Classroom Management and Environment

Domain 3: Knowledge of Subject Matter

Domain 4: Clarity of Presentation

Domain 5: Availability and Helpfulness

Domain 6: Evaluation and Quality of Examination

Domain 7: Student Outcome

Others

Do you think 7 domains have adequate coverage of teaching quality?

○ Yes (1)

○ No (2)

If not, please provide additional domain(s) that should be added

The 5-point Likert scale below is proposed to use as a response scale of the 24 proposed survey items. Please rate how appropriate the response scale is

○ Excellent (1)

○ Very good (2)

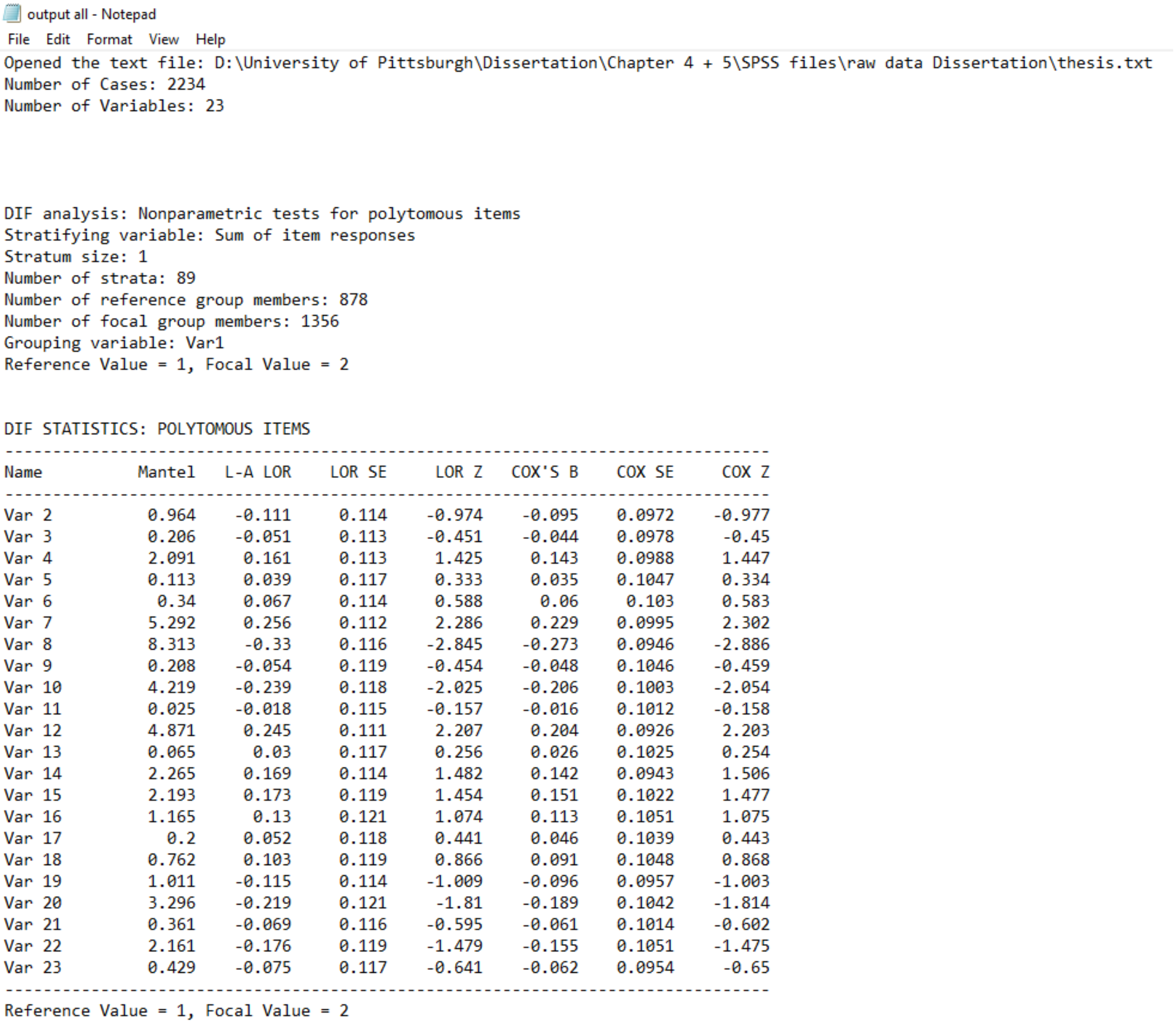
○ Good (3)

○ Fair (4)

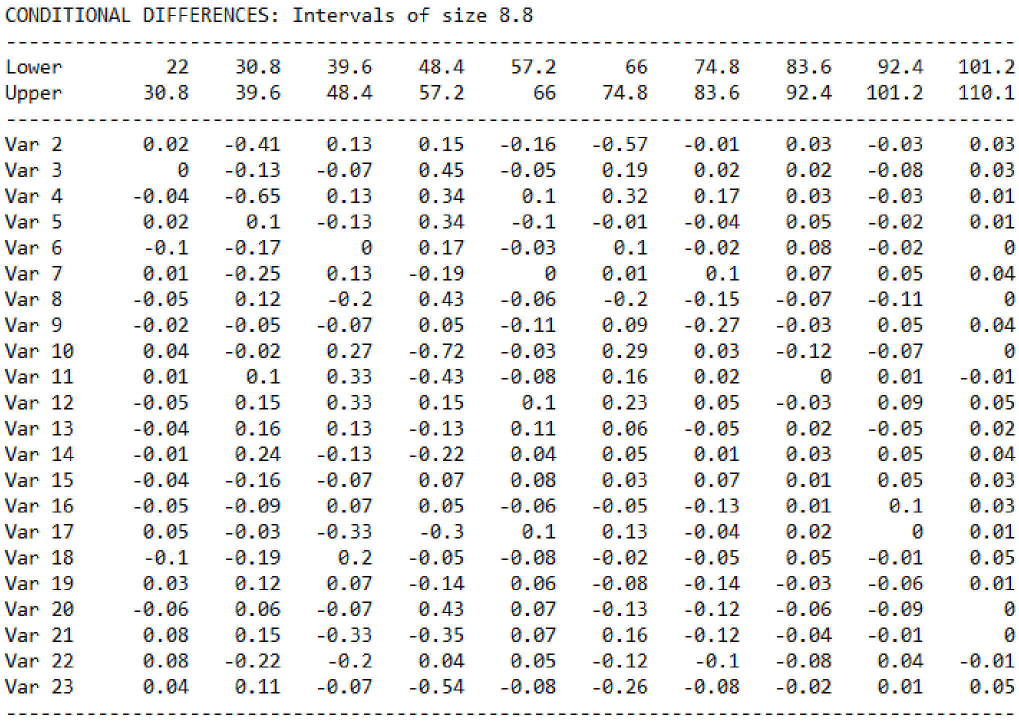
○ Poor (5)

APPENDIX C

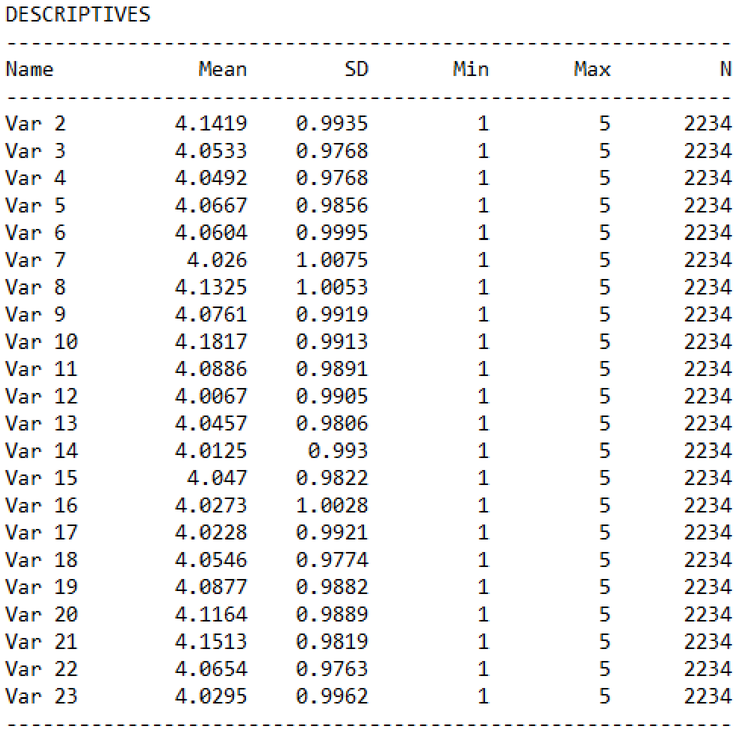
OUTPUT FROM THE DIFAS PROGRAM



```
Opened the text file: D:\University of Pittsburgh\Dissertation\Chapter 4 + 5\SPSS files\raw data Dissertation\thesis.txt
Number of Cases: 2234
Number of Variables: 23




DIF analysis: Nonparametric tests for polytomous items
Stratifying variable: Sum of item responses
Stratum size: 1
Number of strata: 89
Number of reference group members: 878
Number of focal group members: 1356
Grouping variable: Var1
Reference Value = 1, Focal Value = 2


DIF STATISTICS: POLYTOMOUS ITEMS
--------------------------------------------------------------------------------
Name          Mantel   L-A LOR    LOR SE     LOR Z   COX'S B    COX SE     COX Z
--------------------------------------------------------------------------------
Var 2          0.964    -0.111     0.114    -0.974    -0.095    0.0972    -0.977
Var 3          0.206    -0.051     0.113    -0.451    -0.044    0.0978     -0.45
Var 4          2.091     0.161     0.113     1.425     0.143    0.0988     1.447
Var 5          0.113     0.039     0.117     0.333     0.035    0.1047     0.334
Var 6           0.34     0.067     0.114     0.588      0.06     0.103     0.583
Var 7          5.292     0.256     0.112     2.286     0.229    0.0995     2.302
Var 8          8.313     -0.33     0.116    -2.845    -0.273    0.0946    -2.886
Var 9          0.208    -0.054     0.119    -0.454    -0.048    0.1046    -0.459
Var 10         4.219    -0.239     0.118    -2.025    -0.206    0.1003    -2.054
Var 11         0.025    -0.018     0.115    -0.157    -0.016    0.1012    -0.158
Var 12         4.871     0.245     0.111     2.207     0.204    0.0926     2.203
Var 13         0.065      0.03     0.117     0.256     0.026    0.1025     0.254
Var 14         2.265     0.169     0.114     1.482     0.142    0.0943     1.506
Var 15         2.193     0.173     0.119     1.454     0.151    0.1022     1.477
Var 16         1.165      0.13     0.121     1.074     0.113    0.1051     1.075
Var 17           0.2     0.052     0.118     0.441     0.046    0.1039     0.443
Var 18         0.762     0.103     0.119     0.866     0.091    0.1048     0.868
Var 19         1.011    -0.115     0.114    -1.009    -0.096    0.0957    -1.003
Var 20         3.296    -0.219     0.121     -1.81    -0.189    0.1042    -1.814
Var 21         0.361    -0.069     0.116    -0.595    -0.061    0.1014    -0.602
Var 22         2.161    -0.176     0.119    -1.479    -0.155    0.1051    -1.475
Var 23         0.429    -0.075     0.117    -0.641    -0.062    0.0954     -0.65
--------------------------------------------------------------------------------
Reference Value = 1, Focal Value = 2
```



```
CONDITIONAL DIFFERENCES: Intervals of size 8.8
---------------------------------------------------------------------------------------
Lower         22    30.8    39.6    48.4    57.2      66    74.8    83.6    92.4   101.2
Upper       30.8    39.6    48.4    57.2      66    74.8    83.6    92.4   101.2   110.1
---------------------------------------------------------------------------------------
Var 2       0.02   -0.41    0.13    0.15   -0.16   -0.57   -0.01    0.03   -0.03    0.03
Var 3          0   -0.13   -0.07    0.45   -0.05    0.19    0.02    0.02   -0.08    0.03
Var 4      -0.04   -0.65    0.13    0.34     0.1    0.32    0.17    0.03   -0.03    0.01
Var 5       0.02     0.1   -0.13    0.34    -0.1   -0.01   -0.04    0.05   -0.02    0.01
Var 6       -0.1   -0.17       0    0.17   -0.03     0.1   -0.02    0.08   -0.02       0
Var 7       0.01   -0.25    0.13   -0.19       0    0.01     0.1    0.07    0.05    0.04
Var 8      -0.05    0.12    -0.2    0.43   -0.06    -0.2   -0.15   -0.07   -0.11       0
Var 9      -0.02   -0.05   -0.07    0.05   -0.11    0.09   -0.27   -0.03    0.05    0.04
Var 10      0.04   -0.02    0.27   -0.72   -0.03    0.29    0.03   -0.12   -0.07       0
Var 11      0.01     0.1    0.33   -0.43   -0.08    0.16    0.02       0    0.01   -0.01
Var 12     -0.05    0.15    0.33    0.15     0.1    0.23    0.05   -0.03    0.09    0.05
Var 13     -0.04    0.16    0.13   -0.13    0.11    0.06   -0.05    0.02   -0.05    0.02
Var 14     -0.01    0.24   -0.13   -0.22    0.04    0.05    0.01    0.03    0.05    0.04
Var 15     -0.04   -0.16   -0.07    0.07    0.08    0.03    0.07    0.01    0.05    0.03
Var 16     -0.05   -0.09    0.07    0.05   -0.06   -0.05   -0.13    0.01     0.1    0.03
Var 17      0.05   -0.03   -0.33    -0.3     0.1    0.13   -0.04    0.02       0    0.01
Var 18      -0.1   -0.19     0.2   -0.05   -0.08   -0.02   -0.05    0.05   -0.01    0.05
Var 19      0.03    0.12    0.07   -0.14    0.06   -0.08   -0.14   -0.03   -0.06    0.01
Var 20     -0.06    0.06   -0.07    0.43    0.07   -0.13   -0.12   -0.06   -0.09       0
Var 21      0.08    0.15   -0.33   -0.35    0.07    0.16   -0.12   -0.04   -0.01       0
Var 22      0.08   -0.22    -0.2    0.04    0.05   -0.12    -0.1   -0.08    0.04   -0.01
Var 23      0.04    0.11   -0.07   -0.54   -0.08   -0.26   -0.08   -0.02    0.01    0.05
---------------------------------------------------------------------------------------
```



```
DESCRIPTIVES
--------------------------------------------------------------
Name              Mean        SD       Min       Max         N
--------------------------------------------------------------
Var 2           4.1419    0.9935         1         5      2234
Var 3           4.0533    0.9768         1         5      2234
Var 4           4.0492    0.9768         1         5      2234
Var 5           4.0667    0.9856         1         5      2234
Var 6           4.0604    0.9995         1         5      2234
Var 7            4.026    1.0075         1         5      2234
Var 8           4.1325    1.0053         1         5      2234
Var 9           4.0761    0.9919         1         5      2234
Var 10          4.1817    0.9913         1         5      2234
Var 11          4.0886    0.9891         1         5      2234
Var 12          4.0067    0.9905         1         5      2234
Var 13          4.0457    0.9806         1         5      2234
Var 14          4.0125     0.993         1         5      2234
Var 15           4.047    0.9822         1         5      2234
Var 16          4.0273    1.0028         1         5      2234
Var 17          4.0228    0.9921         1         5      2234
Var 18          4.0546    0.9774         1         5      2234
Var 19          4.0877    0.9882         1         5      2234
Var 20          4.1164    0.9889         1         5      2234
Var 21          4.1513    0.9819         1         5      2234
Var 22          4.0654    0.9763         1         5      2234
Var 23          4.0295    0.9962         1         5      2234
--------------------------------------------------------------
```

APPENDIX D

OUTPUT FROM THE GMHDIF PROGRAM

GMHDIF Beta Version

GMH statistic = QMH1

Alpha level = 0.05

Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE 2

Stage 1: QMH = 5.0398 df = 4 p = 0.2832
Stage 2: QMH = 4.7830 df = 4 p = 0.3103

VARIABLE 3

Stage 1: QMH = 2.6209 df = 4 p = 0.6231
Stage 2: QMH = 2.9625 df = 4 p = 0.5641

VARIABLE 4

Stage 1: QMH = 7.7799 df = 4 p = 0.1000
Stage 2: QMH = 6.7378 df = 4 p = 0.1504

VARIABLE 5

Stage 1: QMH = 2.3965 df = 4 p = 0.6633
Stage 2: QMH = 1.5654 df = 4 p = 0.8150

VARIABLE 6

Stage 1: QMH = 2.9083 df = 4 p = 0.5733
Stage 2: QMH = 3.6641 df = 4 p = 0.4534

VARIABLE 7

Stage 1: QMH = 9.3266 df = 4 p = 0.0534

Stage 2: QMH = 7.4266 df = 4 p = 0.1150

VARIABLE 8

Stage 1: QMH = 9.2689 df = 4 p = 0.0547

Stage 2: QMH = 9.9338 df = 4 p = 0.0416*

VARIABLE 9

Stage 1: QMH = 3.0836 df = 4 p = 0.5439

Stage 2: QMH = 3.3237 df = 4 p = 0.5052

VARIABLE 10

Stage 1: QMH = 5.0803 df = 4 p = 0.2792

Stage 2: QMH = 6.3646 df = 4 p = 0.1735

VARIABLE 11

Stage 1: QMH = 1.1057 df = 4 p = 0.8934

Stage 2: QMH = 0.7669 df = 4 p = 0.9428

VARIABLE 12

Stage 1: QMH = 5.6882 df = 4 p = 0.2237

Stage 2: QMH = 6.6395 df = 4 p = 0.1562

VARIABLE 13

Stage 1: QMH = 1.4144 df = 4 p = 0.8417

Stage 2: QMH = 1.4027 df = 4 p = 0.8437

VARIABLE 14

Stage 1: QMH = 5.2835 df = 4 p = 0.2594

Stage 2: QMH = 5.4250 df = 4 p = 0.2464

VARIABLE 15

Stage 1: QMH = 7.2551 df = 4 p = 0.1230
Stage 2: QMH = 7.8123 df = 4 p = 0.0987

VARIABLE 16

Stage 1: QMH = 9.1073 df = 4 p = 0.0585
Stage 2: QMH = 7.7438 df = 4 p = 0.1014

VARIABLE 17

Stage 1: QMH = 1.7683 df = 4 p = 0.7783
Stage 2: QMH = 1.8969 df = 4 p = 0.7547

VARIABLE 18

Stage 1: QMH = 7.9892 df = 4 p = 0.0920
Stage 2: QMH = 8.0841 df = 4 p = 0.0885

VARIABLE 19

Stage 1: QMH = 8.5597 df = 4 p = 0.0731
Stage 2: QMH = 7.7331 df = 4 p = 0.1019

VARIABLE 20

Stage 1: QMH = 5.5157 df = 4 p = 0.2384
Stage 2: QMH = 6.3136 df = 4 p = 0.1769

VARIABLE 21

Stage 1: QMH = 1.5580 df = 4 p = 0.8163
Stage 2: QMH = 3.3533 df = 4 p = 0.5005

VARIABLE 22

Stage 1: QMH = 6.5121 df = 4 p = 0.1640
Stage 2: QMH = 3.8012 df = 4 p = 0.4336

VARIABLE 23

Stage 1: QMH = 10.0340 df = 4 p = 0.0399*
Stage 2: QMH = 10.0340 df = 4 p = 0.0399*

07:09 PM, Thursday 21-June-2018

Analyses time

Start = 7:09:00 PM
End = 7:09:01 PM

1 seconds

BIBLIOGRAPHY

Abrami, P. C. (1985). Dimensions of Effective College Instruction. *The Review of Higher Education, 8*(3), 211-228. doi: 10.1353/rhe.1985.0018

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of Student Ratings of Instruction: What We Know and What We Do Not. *Journal of educational psychology, 82*(2), 219. doi: 10.1037/0022-0663.82.2.219

Agresti, A. (2013). *Categorical data analysis* (3rd;3; ed. Vol. 792). Hoboken, NJ: Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81. doi: 10.1007/bf02293746

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi: 10.1007/bf02293814

Angoff, W. H. (1993). *Perspective on differential item functioning methodology*: In P.W.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education, 30*(6), 723-748. doi: 10.1080/03075070500340101

Armstrong, T. S., Cohen, M. Z., Eriksen, L., & Cleeland, C. (2005). Content validity of self-report measurement instruments: An illustration from the development of the Brain Tumor Module of the M.D. Anderson Symptom Inventory. *Oncology Nursing Forum, 32*(3), 669-676. doi: 10.1188/05.ONF.669-676

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: a guide to designing, building, and operating large-scale faculty evaluation systems* (3rd ed.). Bolton, Mass: Anker Pub. Co.

ASEAN University Network. (2011). *ASEAN University Network - Quality Assurance Guidelines*. Bangkok, Thailand.

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. doi: 10.1080/10705510903008204

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *Journal of Management, 41*(6), 1561-1577. doi: 10.1177/0149206315591075

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21*, 1105-1106.

Barnes, D. C., Engelland, B. T., Matherine, C. F., Martin, W. C., Orgeron, C. P., Ring, J. K., . . . Williams, Z. (2008). Developing a Psychometrically Sound Measure of Collegiate Teaching Proficiency. *College Student Journal*, *42*(1), 199.

Barth, M. M. (2008). Deciphering Student Evaluations of Teaching: A Factor Analysis Approach. *Journal of Education for Business, 84*(1), 40-46. doi: 10.3200/JOEB.84.1.40-46

Basow, S. A., & Montgomery, S. (2005). Student Ratings and Professor Self-Ratings of College Teaching: Effects of Gender and Divisional Affiliation. *Journal of Personnel Evaluation in Education, 18*(2), 91-106. doi: 10.1007/s11092-006-9001-8

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An Argument Approach to Observation Protocol Validity. *Educational Assessment, 17*(2-3), 62-87. doi: 10.1080/10627197.2012.715014

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: how much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*(6), 593-601. doi: 10.1080/02602930500260688

Blaikie, N. (2003). *Analyzing quantitative data: from description to explanation*. Thousand Oaks, CA: Sage Publications Ltd.

Bollen, K. A., & Pearl, J. (2012). Eight myths about causality and structural equation models. In S. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301-330). New York, NY: Springer.

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF Detection Methods. *Applied Measurement in Education, 15*(2), 113-141. doi: 10.1207/S15324818AME1502_01

Booth, T., & Hughes, D. J. (2014). Exploratory Structural Equation Modeling of Personality Data. *Assessment, 21*(3), 260-271. doi: 10.1177/1073191114528029

Brandenburg, D. C., & Slinde, J. A. (1977). Student Ratings of Instruction: Validity and Normative Interpretations. *Research in Higher Education, 7*(1), 67-78. doi: 10.1007/BF00991945

Braun, E., & Leidner, B. (2009). Academic Course Evaluation: Theoretical and Empirical Distinctions Between Self-Rated Gain in Competences and Satisfaction with Teaching Behavior. *European Psychologist*, *14*(4), 297-306. doi: 10.1027/1016-9040.14.4.297

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second ed.). London;New York;: The Guilford Press.

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education, 33*(5), 567-576. doi: 10.1080/02602930701699049

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks [Calif.]: Sage Publications.

Capa-Aydin, Y. (2016). Student evaluation of instruction: comparison between in-class and online methods. *Assessment & Evaluation in Higher Education, 41*(1), 112-126. doi: 10.1080/02602938.2014.987106

Carifio, J., & Perla, R. J. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences, 3*(3), 106-116. doi: 10.3844/jssp.2007.106.116

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of educational psychology, 84*(4), 563-572. doi: http://dx.doi.org/10.1037/0022-0663.84.4.563

Centra, J. A., & Gaubatz, N. B. (2000). Is There Gender Bias in Student Evaluations of Teaching? *The Journal of Higher Education, 71*(1), 17-33.

Chatterjee, A., Ghosh, C., & Bandyopadhyay, S. (2009). Assessing students' rating in higher education: A SERVQUAL approach. *Total Quality Management, 20*(10), 1095-1109. doi: 10.1080/14783360903247114

Cheung, D. (2000). Evidence of a Single Second-Order Factor in Student Ratings of Teaching Effectiveness. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(3), 442-460. doi: 10.1207/S15328007SEM0703_5

Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability : A Monte Carlo Investigation. *Applied Psychological Measurement, 9*(1), 31-36. doi: 10.1177/014662168500900103

Clauser, B. E. (1993). The Effects of Purification of the Matching Criterion on the Identification of DIF Using the Mantel-Haenszel Procedure. *Applied Measurement in Education, 6*(4), 269-279. doi: 10.1207/s15324818ame0604_2

Clauser, B. E., Margolis, M. J., Holtman, M. C., Katsufrakis, P. J., & Hawkins, R. E. (2012). Validity considerations in the assessment of professionalism. *Advances in Health Sciences Education, 17*(2), 165-181. doi: 10.1007/s10459-010-9219-6

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1991). Influence of the Criterion Variable on the Identification of Differentially Functioning Test Items Using the Mantel-Haenszel Statistic. *Applied Psychological Measurement, 15*(4), 353-359. doi: 10.1177/014662169101500405

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures To Identify Differentially Functioning Test Items. An NCME Instructional Module. *Educational Measurement: Issues and Practice, 17*(1), 31.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1994). The Effects of Score Group Width on the Mantel-Haenszel Procedure. *Journal of Educational Measurement, 31*(1), 67-78. doi: 10.1111/j.1745-3984.1994.tb00435.x

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A Comparison of Alternative Matching Strategies for DIF Detection in Tests That Are Multidimensional. *Journal of Educational Measurement, 33*(2), 202-214. doi: 10.1111/j.1745-3984.1996.tb00489.x

Clayson, D. E. (2009). Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature. *Journal of Marketing Education, 31*(1), 16-30. doi: 10.1177/0273475308324086

Coffey, M., & Gibbs, G. (2001). The Evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK Higher Education. *Assessment & Evaluation in Higher Education, 26*(1), 89-93. doi: 10.1080/02602930020022318

Cohen, E. H. (2005). Student evaluations of course and teacher: factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education, 30*(2), 123-136. doi: 10.1080/0260293042000264235

Cohen, P. A. (1980). Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-Analysis of Findings. *Research in Higher Education, 13*(4), 321-341. doi: 10.1007/BF00976252

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies. *Review of Educational Research, 51*(3), 281-309. doi: 10.2307/1170209

Cohen, P. A. (1982). Validity of Student Ratings in Psychology Courses: A Research Synthesis. *Teaching of Psychology, 9*(2), 78-82. doi: 10.1207/s15328023top0902_3

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, N.J: L. Erlbaum Associates.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student Ratings of College Teaching: Reliability, Validity, and Usefulness. *Review of Educational Research, 41*(5), 511-535. doi: 10.2307/1169890

Crumbley, D. L., Flinn, R. E., & Reichelt, K. J. (2010). What is Ethical About Grade Inflation and Coursework Deflation? *Journal of Academic Ethics, 8*(3), 187-197. doi: 10.1007/s10805-010-9117-9

d'Apollonia, S., & Abrami, P. C. (1997). Navigating Student Ratings of Instruction. *American Psychologist, 52*(11), 1198-1208. doi: 10.1037/0003-066X.52.11.1198

Deng, L., Marcoulides, G. A., & Yuan, K.-H. (2015). Psychometric Properties of Measures of Team Diversity With Likert Data. *Educational and Psychological Measurement, 75*(3), 512-534. doi: 10.1177/0013164414541275

Díaz, S. R., Swan, K., Ice, P., & Kupczynski, L. (2010). Student ratings of the importance of survey items, multiplicative factor analysis, and the validity of the community of inquiry survey. *The Internet and Higher Education, 13*(1–2), 22-30. doi: http://dx.doi.org/10.1016/j.iheduc.2009.11.004

Eiszler, C. F. (2002). College Students' Evaluations of Teaching and Grade Inflation. *Research in Higher Education, 43*(4), 483-501. doi: 10.1023/A:1015579817194

Feldman, K. A. (1978). Course Characteristics and College Students' Ratings of Their Teachers: What We Know and What We Don't. *Research in Higher Education, 9*(3), 199-242. doi: 10.1007/BF00976997

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-143): Springer.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *The Phi Delta Kappan, 94*(3), 24-28. doi: 10.2307/41763671

Fidalgo, Á. M. (2011). GMHDIF: A Computer Program for Detecting DIF in Dichotomous and Polytomous Items Using Generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement, 35*(3), 247-249. doi: 10.1177/0146621610375691

Fidalgo, Á. M. (2011). A new approach for differential item functioning detection using Mantel-Haenszel methods. The GMHDIF program. *The Spanish journal of psychology, 14*(2), 1018-1022. doi: 10.5209/rev_SJOP.2011.v14.n2.47

Fidalgo, Á. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel Methods for Differential Item Functioning Detection. *Educational and Psychological Measurement, 68*(6), 940-958. doi: 10.1177/0013164408315265

Fidalgo, Á. M., & Scalon, J. D. (2010). Using Generalized Mantel-Haenszel Statistics to Assess DIF Among Multiple Groups. *Journal of Psychoeducational Assessment, 28*(1), 60-69. doi: 10.1177/0734282909337302

Fink, L. D. (2013). *Creating significant learning experiences: an integrated approach to designing college courses* (Revis and updat;2; ed.). San Francisco: Jossey-Bass.

Floyd, F. J., & Widaman, K. F. (1995). Factor Analysis in the Development and Refinement of Clinical Assessment Instruments. *Psychological Assessment, 7*(3), 286-299. doi: 10.1037/1040-3590.7.3.286

Fresko, B., & Nasser, F. (2001). Interpreting student ratings: consultation, instructional modification, and attitudes towards course evaluation. *Studies in Educational Evaluation, 27*(4), 291-305. doi: http://dx.doi.org/10.1016/S0191-491X(01)00031-1

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Research in Higher Education, 53*(3), 353-374. doi: 10.1007/s11162-011-9229-0

Gandek, B., Ware Jr, J. E., Aaronson, N. K., Alonso, J., Apolone, G., Bjorner, J., . . . Sullivan, M. (1998). Tests of data quality, scaling assumptions, and reliability of the SF- 36 in eleven countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology, 51*(11), 1149-1158. doi: 10.1016/S0895-4356(98)00106-1

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: the perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603-615. doi: 10.1080/03075070701573773

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., Whitehurst, G. J. R., & Croft, M. (2011). Passing Muster: Evaluating Teacher Evaluation Systems: Brookings

Institution U6 - ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info:sid/summon.serialssolutions.com&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.atitle=Passing+Muster%3A+Evaluating+Teacher+Evaluation+Systems&rft.au=Glazerman%2C+Steven&rft.au=Goldhaber%2C+Dan&rft.au=Loeb%2C+Susanna&rft.au=Raudenbush%2C+Stephen&rft.date=2011-04-26&rft.pub=Brookings+Institution&rft.externalDocID=2011042600136314&paramdict=en-US U7 - Paper.

Goldstein, G. S., & Benassi, V. A. (2006). Students' and Instructors' Beliefs about Excellent Lecturers and Discussion Leaders. *Research in Higher Education, 47*(6), 685-707. doi: 10.1007/s11162-006-9011-x

Granberg-Rademacker, J. S. (2010). An Algorithm for Converting Ordinal Scale Measurement Data to Interval/Ratio Scale. *Educational and Psychological Measurement*, *70*(1), 74-90. doi: 10.1177/0013164409344532

Greenwald, A. G. (1997). Validity Concerns and Usefulness of Student Ratings of Instruction. *American Psychologist, 52*(11), 1182-1186. doi: 10.1037/0003-066X.52.11.1182

Guilera, G., Gómez-Benito, J., & Hidalgo, M. D. (2009). Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF. *Psicothema, 21*(3), 492-498.

Gursoy, D., & Umbreit, W. T. (2005). Exploring Students’ Evaluations of Teaching Effectiveness: What Factors are Important? *Journal of Hospitality & Tourism Research, 29*(1), 91-109. doi: 10.1177/1096348004268197

Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The Relative Merits of Different Types of Overall Evaluations of Teaching Effectiveness. *Research in Higher Education, 45*(3), 311-323. doi: 10.1023/B:RIHE.0000019592.78752.da

Harwell, M. R., & Gatti, G. G. (2001). Rescaling Ordinal Data to Interval Data in Educational Research. *Review of Educational Research, 71*(1), 105-131. doi: 10.3102/00346543071001105

Hativa, N., Barak, R., & Simhi, E. (2001). Exemplary University Teachers: Knowledge and Beliefs Regarding Effective Teaching Dimensions and Strategies. *The Journal of Higher Education, 72*(6), 699-729. doi: 10.2307/2672900

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London;New York;: Routledge.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment, 7*(3), 238-247. doi: 10.1037/1040-3590.7.3.238

Hill, H. C., Kapitula, L., & Umland, K. (2011). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal, 48*(3), 794-831. doi: 10.3102/0002831210387916

Holland, P. W., & Thayer, D. T. (1986). DIFFERENTIAL ITEM FUNCTIONING AND THE MANTEL-HAENSZEL PROCEDURE. *ETS Research Report Series, 1986*(2), i-24. doi: 10.1002/j.2330-8516.1986.tb00186.x

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). NJ: Erlbaum: Hillsdale.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning: theory and practice*. Hillsdale: Lawrence Erlbaum Associates.

Hu, L.-t., & 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. doi: 10.1080/10705519909540118

James, D. E., Schraw, G., & Kuch, F. (2015). Using the sampling margin of error to assess the interpretative validity of student evaluations of teaching. *Assessment & Evaluation in Higher Education, 40*(8), 1123-1141. doi: 10.1080/02602938.2014.972338

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education, 38*(12), 1217-1218. doi: 10.1111/j.1365-2929.2004.02012.x

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology, 7*(4), 396-403. doi: 10.9734/BJAST/2015/14975

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger Publishers.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (2014). *Designing Teacher Evaluation Systems : New Guidance from the Measures of Effective Teaching Project*. Somerset, NJ, USA: Wiley.

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating Psychology Students' Perceptions of Teachers Using the Teacher Behavior Checklist. *Teaching of Psychology, 37*(1), 16-20. doi: 10.1080/00986280903426282

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor Analysis of Its Utility for Evaluating Teaching. *Teaching of Psychology, 33*(2), 84-91. doi: 10.1207/s15328023top3302_1

Kelly, H. F., Ponton, M. K., & Rovai, A. P. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. *The Internet and Higher Education, 10*(2), 89-101. doi: 10.1016/j.iheduc.2007.02.001

Kember, D., & Doris, Y. P. L. (2011). Disciplinary Differences in Student Ratings of Teaching Quality. *Research in Higher Education, 52*(3), 278-299. doi: 10.1007/s11162-010-9194-z

Kember, D., & Leung, D. Y. P. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education, 33*(4), 341-353. doi: 10.1080/02602930701563070

Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the Use of Student Feedback Questionnaires Improve the Overall Quality of Teaching? *Assessment & Evaluation in Higher Education, 27*(5), 411-425. doi: 10.1080/0260293022000009294

Kember, D., & Wong, A. (2000). Implications for Evaluation from a Study of Students' Perceptions of Good and Poor Teaching. *Higher Education, 40*(1), 69-97. doi: 10.1023/A:1004068500314

Kidder, L., & Judd, C. (1986). *Research methods in social science*. New York, NY: CBS College Publishing.

Kohn, J., & Hatfield, L. (2006). The role of gender in teaching effectiveness ratings of faculty. *Academy of Educational Leadership Journal, 10*(3), 121.

Kolitch, E., & Dean, A. V. (1999). Student Ratings of Instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher Education, 24*(1), 27-42.

Koon, J., & Murray, H. G. (1995). Using Multiple Outcomes to Validate Student Ratings of Overall Teacher Effectiveness. *The Journal of Higher Education, 66*(1), 61-81.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items. *Educational and Psychological Measurement, 65*(6), 935-953. doi: 10.1177/0013164405275668

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). Bingley: Emerald Group.

Kwan, K.-p. (1999). How Fair are Student Ratings in Assessing the Teaching Performance of University Teachers? *Assessment & Evaluation in Higher Education, 24*(2), 181-195. doi: 10.1080/0260293990240207

Linacre, J. M. (2015). A user's guide to Winsteps Ministep Rasch-model computer programs. from http://www.winsteps.com/manuals.htm

Liu, I. M., & Agresti, A. (1996). Mantel-Haenszel-Type Inference for Cumulative Odds Ratios with a Stratified Ordinal Response. *Biometrics, 52*(4), 1223-1234. doi: 10.2307/2532838

Liu, Y., Wu, A. D., & Zumbo, B. D. (2010). The Impact of Outliers on Cronbach's Coefficient Alpha Estimate of Reliability: Ordinal/Rating Scale Item Responses. *Educational and Psychological Measurement, 70*(1), 5-21. doi: 10.1177/0013164409344548

Liu, Y., & Zumbo, B. D. (2007). The Impact of Outliers on Cronbach's Coefficient Alpha Estimate of Reliability: Visual Analogue Scales. *Educational and Psychological Measurement, 67*(4), 620-634. doi: 10.1177/0013164406296976

Magis, D., Béland, S., Tuerlinckx, F., & de Boeck, P. A. L. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862. doi: 10.3758/BRM.42.3.847

Mantel, N. (1963). Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. *Journal of the American Statistical Association, 58*(303), 690-700. doi: 10.1080/01621459.1963.10500879

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Marks, R. B. (2000). Determinants of Student Evaluations of Global Measures of Instructor and Course Value. *Journal of Marketing Education, 22*(2), 108-119. doi: 10.1177/0273475300222005

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319-384). New York: Springer.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*(1), 9-18. doi: 10.1016/0742-051X(91)90054-S

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A New Look at the Big Five Factor Structure Through Exploratory Structural Equation Modeling. *Psychological Assessment, 22*(3), 471-491. doi: 10.1037/a0019227

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology, 10*(1), 85-110. doi: 10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 439-476. doi: 10.1080/10705510903008220

Marsh, H. W., & Roche, L. A. (1997). Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. *American Psychologist, 52*(11), 1187-1197. doi: 10.1037/0003-066X.52.11.1187

Marsh, H. W., & Roche, L. A. (2000). Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders? *Journal of educational psychology, 92*(1), 202-228. doi: 10.1037/0022-0663.92.1.202

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic. *Educational and Psychological Measurement, 52*(2), 443-451. doi: 10.1177/0013164492052002020

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF Analyses: The Effects of Matching on Unidimensional Subtest Scores. *Applied Psychological Measurement, 22*(4), 357-367. doi: 10.1177/014662169802200404

McKeachie, W. J. (1990). Research on College Teaching: The Historical Background. *Journal of educational psychology, 82*(2), 189-200. doi: 10.1037/0022-0663.82.2.189

McKeachie, W. J. (1997). Student Ratings: The Validity of Use. *American Psychologist, 52*(11), 1218-1225. doi: 10.1037/0003-066X.52.11.1218

McKone, K. E. (1999). Analysis of Student Feedback Improves Instructor Effectiveness. *Journal of Management Education, 23*(4), 396-415. doi: 10.1177/105256299902300406

McPherson, M. A., & Jewell, R. T. (2007). Leveling the Playing Field: Should Student Evaluation Scores be Adjusted? *Social Science Quarterly, 88*(3), 868-881. doi: 10.1111/j.1540-6237.2007.00487.x

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores?: A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35*(1), 37-51. doi: 10.1057/palgrave.eej.9050042

Meyer, J. P., Doromal, J. B., Wei, X., & Zhu, S. (2016). A Criterion-Referenced Approach to Student Ratings of Instruction. *Research in Higher Education, 58*(5), 545. doi: 10.1007/s11162-016-9437-8

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A Bifactor Exploratory Structural Equation Modeling Framework for the Identification of Distinct Sources of Construct-

Relevant Psychometric Multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(1), 116-139. doi: 10.1080/10705511.2014.961800

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies, 35*(5), 547-552. doi: 10.1080/03055690902880299

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of educational psychology, 75*(1), 138-149. doi: 10.1037/0022-0663.75.1.138

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. doi: 10.1007/s10459-010-9222-y

Ong Chee Bin, J. (2016). *Guide to AUN-QA assessment at institutional level*. Bangkok, Thailand: ASEAN University Network (AUN).

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*(2), 197-209. doi: 10.1007/s11135-007-9112-4

Onwuegbuzie, A. J., Witcher, A. E., Kathleen, M. T. C., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' Perceptions of Characteristics of Effective College Teachers: A Validity Study of a Teaching Evaluation Form Using a Mixed-Methods Analysis. *American Educational Research Journal, 44*(1), 113-160. doi: 10.3102/0002831206298169

Oon, P.-T., Spencer, B., & Kam, C. C. S. (2017). Psychometric quality of a student evaluation of teaching survey in higher education. *Assessment & Evaluation in Higher Education, 42*(5), 788-713. doi: 10.1080/02602938.2016.1193119

Ory, J. C., & Ryan, K. (2001). How Do Student Ratings Measure Up to a New Validity Framework? *New Directions for Institutional Research, 2001*(109), 27-44. doi: 10.1002/ir.2

Padilla, J. L., Hidalgo, M. D., Benitez, I., & Gomez-Benito, J. (2012). Comparison of Three Software Programs for Evaluating DIF by Means of the Mantel-Haenszel Procedure: EASY-DIF, DIFAS and EZDIF. *Psicologica: International Journal of Methodology and Experimental Psychology, 33*(1), 135.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*(2), 150-151. doi: 10.1177/0146621603260686

Penfield, R. D. (2007). An Approach for Categorizing DIF in Polytomous Items. *Applied Measurement in Education, 20*(3), 335-355. doi: 10.1080/08957340701431435

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti Estimator of the Cumulative Common Odds Ratio to DIF Detection in Polytomous Items. *Journal of Educational Measurement, 40*(4), 353-370. doi: 10.1111/j.1745-3984.2003.tb01151.x

Penny, A. R. (2003). Changing the Agenda for Research into Students' Views about University Teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*(3), 399-411. doi: 10.1080/13562510309396

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. doi: 10.1016/S0001-6918(99)00050-5

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297. doi: 10.1037//1040-3590.12.3.287

Renaud, R. D., & Murray, H. G. (2005). Factorial Validity of Student Ratings of Instruction. *Research in Higher Education, 46*(8), 929-953. doi: 10.2307/40197397

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387-415. doi: 10.1080/02602930500099193

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity Issues in the Likert and Thurstone Approaches to Attitude Measurement. *Educational and Psychological Measurement, 59*(2), 211-233. doi: 10.1177/00131649921969811

Roberts, T. R. M. (2008). *A test of the reliability of student ratings over time.* (Dissertation/Thesis), ProQuest, UMI Dissertations Publishing. Retrieved from http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1LT8MwDLYQSAiBBOMR3sofaOkjfZ0Q2pgmwZEDt6lNUjSpdGwtB_49dtIitLELx7QXp40-2_HnzwBh4HrOCiboRBWlr0KVooMLpIh0FCshNUbnoS9TtUKy7FUhur_dg6RBbjWXdGl-h3m4SGnKzv3HwqEpUlRt7UZqICT7QZzZ2u1fyTtmRQkJi0edBk-_Ttcg2fiZ8SH0lFEjxzRT7z_d0uvyjf-1-gj2R7-K8APY0vUxTW_umB4nED9wDEBbPi85hod8qauZVfP-okeNVcPkdHTqt4YTB5TTiPpTuB0_vgwnTm_SVFXVFAMZITIRoxs7g4OcCPV1axrvFIOdEk-_ZuSRGH4DBruv2fMonTwN7XLQL93GdIe5i5ahAzT7cGI3OQceedILlNSBX4QiDlVeIPBGIswxQiFVsQtgG-y53PjmCvYsYYPuQK5hu11-6hsrS_sNRH60Mw

Ryan, C. H. (2008). *Using hierarchical generalized linear modeling for detection of differential item functioning in a polytomous item response theory framework: An evaluation and comparison with generalized Mantel -Haenszel.* (Dissertation/Thesis), ProQuest Dissertations Publishing. Retrieved from http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwpV1LS8QwEB58XERBRUVdhcF71bYx3XoRUZe9ePO-pElWCjVdaz3s_kZ_lJmkrVplLx5LQwnJdJ7fzAcQR-eXQU8nsDCLQymsQZAxU4lQXCjJEzXkXEeRzHogy66pv7nuVks61a1KSVnzC6rpMSo03sxeA6KRonJrw6mxCuuhNU2Jr97-Fb7buMgaTzZs5j51z-yXUnaWZrQNLRjADWTK1UvXL90b4Piffe_A5v23OvwurGizBx8OPoBEju3KC_b28NlPpc4XWiE5pKJCx51DC627i0rXDstlsJxiS7ZilUaBlBRGsppNxhdzgwJnZTGvS0o3-AWVh-hqdA2Vc5y2WLFrvDX4NYochVEoO85EpPTxj809koQUGIyFjcsXutiHs9HD0904aA9n0vw1b5PuZOID2BKE8De16wRUh4BWVqwCCpnkVoqE5CJLlfUt4vRKRank8REMlnzxeOnbAWx4HAilVk5gra7e9amfdvsJLwviOQ

Safavi, S. A., Bakar, K. A., Tarmizi, R. A., & Alwi, N. H. (2012). What do higher education instructors consider useful regarding student ratings of instruction? Limitations and recommendations. *Procedia - Social and Behavioral Sciences, 31*, 653-657. doi: http://dx.doi.org/10.1016/j.sbspro.2011.12.119

Santhanam, E., & Hicks, O. (2002). Disciplinary, Gender and Course Year Influences on Student Perceptions of Teaching: Explorations and implications. *Teaching in Higher Education, 7*(1), 17-31. doi: 10.1080/13562510120100364

Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction, 16*(5), 401-415. doi: http://dx.doi.org/10.1016/j.learninstruc.2006.09.002

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The Validity of Student Evaluation of Teaching in Higher Education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405. doi: 10.1080/713611436

Simendinger, E., El-Kassar, A.-N., Gonzalez-Perez, M. A., Crawford, J., Thomason, S., Reynet, P., . . . Edwards, J. (2017). Teaching effectiveness attributes in business schools. *International Journal of Educational Management, 31*(6), 780-800. doi: 10.1108/IJEM-05-2016-0108

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The Influence of Student Sex and Instructor Sex on Student Ratings of Instructors: Results from a College of Communication. *Women's Studies in Communication, 30*(1), 64-77. doi: 10.1080/07491409.2007.10162505

Somes, G. W. (1986). The Generalized Mantel-Haenszel Statistic. *The American Statistician, 40*(2), 106-108. doi: 10.1080/00031305.1986.10475369

Sorenson, D. L., & Johnson, T. D. (2003). *Online student ratings of instruction* (Vol. no. 96). San Francisco, Calif: Jossey-Bass.

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation, 36*(4), 121-131. doi: 10.1016/j.stueduc.2011.02.001

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research, 83*(4), 598-642. doi: 10.3102/0034654313496870

Spooren, P., Mortelmans, D., & Thijssen, P. (2012). 'Content' versus 'style': acquiescence in student evaluation of teaching? *British Educational Research Journal, 38*(1), 3-21. doi: 10.1080/01411926.2010.523453

Spray, J., & Miller, T. (1994). [Identifying nonuniform DIF in polytomously scored test items (American College Testing Research Report Series 94-1)].

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function Analysis Methods in Detecting Differential Item Functioning for Polytomous Items. *Applied Measurement in Education, 18*(4), 313-350. doi: 10.1207/s15324818ame1804_1

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*(4), 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Thrush, C. R., Putten, J. V., Rapp, C. G., Pearson, L. C., Berry, K. S., & O'Sullivan, P. S. (2007). Content Validation of the Organizational Climate for Research Integrity (OCRI) Survey. *Journal of Empirical Research on Human Research Ethics: An International Journal, 2*(4), 35-52. doi: 10.1525/jer.2007.2.4.35

Tian, F. (1999). *Detecting differential item functioning in polytomous items.* (Doctoral), University of Ottawa.

Toland, M. D., & De Ayala, R. J. (2005). A Multilevel Factor Analysis of Students’ Evaluations of Teaching. *Educational and Psychological Measurement, 65*(2), 272-296. doi: 10.1177/0013164404268667

Tucker, B., Oliver, B., & Gupta, R. (2013). Validating a teaching survey which drives increased response rates in a unit survey. *Teaching in Higher Education, 18*(4), 427-439. doi: 10.1080/13562517.2012.725224

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. doi: 10.1016/j.stueduc.2016.08.007

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness: a brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211. doi: 10.1080/0260293980230207

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological Distance Between Categories in the Likert Scale: Comparing Different Numbers of Options. *Educational and Psychological Measurement, 72*(4), 533-546. doi: 10.1177/0013164411431162

Wang, W.-C., & Su, Y.-H. (2004). Factors Influencing the Mantel and Generalized Mantel-Haenszel Methods for the Assessment of Differential Item Functioning in Polytomous Items. *Applied Psychological Measurement, 28*(6), 450-480. doi: 10.1177/0146621604269792

Weijters, B. (2006). *Rsponse styles in consumer research.* (Doctoral), Ghent Univeristy.

Wigington, H., Tollefson, N., & Rodriguez, E. (1989). Students' Ratings of Instructors Revisited: Interactions among Class and Instructor Variables. *Research in Higher Education, 30*(3), 331-344. doi: 10.1007/BF00992608

Wolf, R. (1967). Evaluation of Several Formulae for Correction of Item-Total Correlations in Item Analysis. *Journal of Educational Measurement, 4*(1), 21-26. doi: 10.1111/j.1745-3984.1967.tb00565.x

Worthington, A. C. (2002). The Impact of Student Perceptions and Characteristics on Teaching Evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education, 27*(1), 49-64. doi: 10.1080/02602930120105054

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two Quantitative Approaches for Estimating Content Validity. *Western Journal of Nursing Research, 25*(5), 508-518. doi: 10.1177/0193945903252998

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*(1), 55-76. doi: 10.1080/13562510601102131

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21-29.

Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 505-509). Thousand Oaks, CA: Sage.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement, 30*(3), 233-251. doi: 10.1111/j.1745-3984.1993.tb00425.x