The Pennsylvania State University

The Graduate School

College of Education

DEVELOPMENT AND EVALUATION OF AUTHENTIC RUBRIC

A Dissertation in

Educational Psychology

by

Wik Hung Pun

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2017

The dissertation of Wik Hung Pun was reviewed and approved* by the following:

Hoi K. Suen
Distinguished Professor of Educational Psychology
Dissertation Adviser
Chair of Committee

Pui-Wa Lei
Professor of Education (Educational Psychology)

Rayne A. Sperling
Associate Professor of Education (Educational Psychology)

Dennis K.J. Lin
Distinguished Professor of Statistics and Supply Chain

Peggy N. Van Meter
Associate Professor of Education (Educational Psychology)
Professor in Charge

*Signatures are on file in the Graduate School

ABSTRACT

Systematic evaluation is a key component of reliable and valid performance assessment. Scoring rubrics are often used to achieve this goal by promoting inter-rater agreement evaluating performances (Johnson, Penny, & Gordon, 2009). More recently, some scholars (e.g., Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Goffin & Olson, 2011; Pollitt, 2004, 2012) advocated for the use of comparative judgment, a scoring procedure adopted from the Thurstone's (Thurstone, 1927a) Law of Comparative Judgment, to replace scoring rubrics in performance assessment. It was argued that the comparative judgment method holds multiple advantages over scoring rubrics. Empirically, some studies were able to show that evaluations elicited via comparative judgment methods had higher criterion-related validity evidence (Goffin et al., 1996; Goffin, Jelley, Powell, & Johnston, 2009; McMahon & Jones, 2014; Olson, Goffin, & Haynes, 2007; Shah, Bradley, Parekh, Wainwright, & Ramchandran, 2013). However, comparative judgment in its original implementation has major drawbacks including laborious evaluation process, not able to discern the absolute quality of the performance, and unable to communicate the evaluation standards to examinee effectively. In this study, a new implementation of the comparative judgment method termed authentic rubric was proposed. The authentic rubric replaced the scoring categories in the scoring rubric with expert evaluated performances and asked the raters to compare performances to be evaluated against these anchors.

22 raters were recruited to evaluate 100 argumentative essays using either a holistic rubric or the proposed authentic rubric. The authentic rubrics were constructed by selecting essays evaluated by two professional raters. Five hypotheses related to user experience, psychometric properties of the evaluations, and efficiency of the evaluation process were proposed and tested. Among the five hypotheses, only one was confirmed and showed that raters

who used authentic rubric found the evaluation experience to be more enjoyable. Nonetheless, examining the data showed that authentic rubric evaluations had marginally higher reliability and criterion-related validity. Post-hoc analysis revealed that a larger sample size may be needed to reach statistical significance conclusion. Implications of the study findings and future areas of research were addressed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deep appreciation and gratitude to my advisor, Dr. Hoi K. Suen, for the guidance and mentorship he provided me throughout my time in the doctoral program. Taking a class with Dr. Suen during my senior year was the reason I embarked on this journey. I would not be here today without Dr. Suen.

I would also like to thank my committee members for their support. In particular, Dr. Puiwa Lei and Dr. Rayne A. Sperling supported and guided me throughout the years. I have learned a tremendous amount from them during my years of study at the Pennsylvania State University.

Finally, I would like to thank my family and friends for their patience and support. Pursuing the doctoral degree was much tougher and longer than I anticipated. I was only able to overcome the challenges during all these years because of them.

Chapter 1

1.1 BACKGROUND

Performance assessment, also referred to as direct assessment (Kane, Crooks, & Cohen, 1999), authentic assessments, alternative assessments, and performance-based assessment (Baker, O'Neil, & Linn, 1993), requires the examinees to demonstrate their domain knowledge and skills in an open-ended settings (Baker et al., 1993; Johnson et al., 2009). A performance can be conceptualized as consisting four components: 1) purpose of the assessment; 2) task that elicits the performance; 3) a response format that focuses the examinee's performance; and 4) systematic methods for evaluating the performances (Klein et al., 1998; Ruiz-Primo & Shavelson, 1996; Stiggins, 1987). The last component of performance assessment is particularly important since the scoring process requires raters' subjective judgment. Without systematic and structured evaluation process, there is no guarantee that raters would reach consensus which can lead to different scores being awarded to the same response (Livingston, 2009).

In fact, early research found that, without explicit guidance, rater judgments are prone to variability (White, 1985). In response, scoring guides were developed in assessment programs to promote scoring consensus by informing the raters on the criteria being judged and the expected performance levels. Among different scoring guides proposed in the literature, one of the most often used scoring guides is the scoring rubric. As a tool for scoring performances, it describes the level of performance in dimensions that are considered important by content experts (Clauser, 2000). The use of scoring rubrics is claimed to have a positive effect on the consistency of judgment across raters in performance assessment (Johnson et al., 2009). Combining the use of scoring rubric and proper training, agreement among raters can be improved substantially.

However, it is important to point out that substantial degree of variations across and within raters can still be present.

More recently, some scholars advocated for the use of a new scoring procedure adopted from the Thurstone's (1927a) scaling procedure, termed comparative judgment or ordinal comparisons (Attali, 2014; Pollitt, 2004, 2012; Raman & Joachims, 2014; Shah et al., 2013). The comparative judgment method, instead of asking the raters to assign a score to each performance, requires the raters to directly compare the quality of the performances against each other. With sufficient number of comparisons, a rank order of the performances can be obtained, which, in turn can be used to estimate the score of each performance through measurement models. Although the theoretical foundation of comparative judgment was developed by Thurstone (1927a) in the early 1900s, it was not implemented as a scoring procedure in educational assessment until recent years (e.g., Attali, 2014; Heldsinger & Humphry, 2010; Jonesa & Alcocka, 2013; Jones, Swan, & Pollitt, 2014; Kimbell, 2012; Pollitt, 2004). In other disciplines including economics (Böckenholt, 2006), medicine (Maydeu-Olivares & Böckenholt, 2008), social psychology (Mussweiler, 2003), personality, and industrial-organizational psychology (Goffin & Olson, 2011), various forms of comparative judgment method were also implemented in recent years.

1.2 STATEMENT OF THE PROBLEM

Proponents of the comparative judgment method argued that comparisons are inherent in all evaluations (Goffin & Olson, 2011; Mussweiler, 2003). Raters do not make an evaluation in a vacuum. Rather, raters would implicitly compare the target with other similar objects to draw evaluation conclusions. By explicitly prompt comparison decisions and model the choices, comparative judgment method can 1) more closely approximate the cognitive processes involved in evaluations, 2) align the implicit standards used by each rater (Böckenholt, 2004; Goffin & Olson, 2011), 3) free raters from the burden of deciding the distance between score intervals (Attali, 2014), 4) eliminate systematic rater variations (Böckenholt, 2004; Bramley, 2007; Shah et al., 2013), and 5) increase the number of response categories beyond common length of Likert-type items (Böckenholt, 2004).

Despite the fact that the comparative judgment method was argued to have theoretical advantages over the use of rating scales and scoring rubrics, there are three concerns associated with the comparative judgment method that limit its utility value in educational assessment. First, the original form of comparative judgment as proposed by Thurstone (1927a), namely pairwise comparison, is labor-intensive (Attali, 2014; Pollitt, 2012). When holistic scoring rubrics are used, raters are responsible for assigning one score to each performance. In other words, raters are only required to evaluate each performance once. In the contrary, the comparative judgment method requires the raters to engage in pairwise comparison with all the performances. Each performance needs to be compared with every other submitted responses. In situations where the number of submissions to be evaluated is large, it is both inefficient and infeasible for the raters to perform a complete pairwise comparison procedure. For example, with 100 performances being evaluated, each rater will have to carry out 4950 pairwise comparisons. Several modifications to the comparative judgment method (e.g., Attali, 2014; Pollitt, 2012) were proposed to alleviate the rater's workload. Although these proposed methods can reduce the required number of comparisons for each rater, it is unclear if these methods can scale with increasing number of performances being evaluated.

Second, while the comparative judgment method can generate a rank order of the performances, the rank order does not contain information regarding the absolute quality of the performances (Böckenholt, 2004). Being judged as the “best” or “worst” performance in the set does not necessarily indicate that the performance is of high or low quality. It is possible that the best performance in the set would still be considered to be inadequate. The lack of information on the absolute quality of the submissions can present a problem in criterion-referenced decisions where the purpose is not to identify which submission is better; instead, the goal is to determine whether the submissions reach certain performance level. Some solutions were proposed to solve this issue, including combining information obtained from ratings and comparative judgment (Böckenholt, 2004) and determine the absolute performance level by examining the ranked performance with post hoc decisions (Jonesa & Alcocka, 2013).

Third, the comparative judgment method cannot inform the respondents on the evaluation criteria effectively. In performance assessment, scoring guides are not only used to facilitate the evaluation process, it also serves the purpose of communicating the expected level of performance at each score level. Both students and teachers value the transparency afforded by the use of scoring guides in performance assessment (Bissell & Lemons, 2006; Schamber & Mahoney, 2006). Additionally, the evaluation criteria specified in scoring guides can be used as pedagogical tools in formative assessments to inform students how they may improve their performances. Comparative judgment method, on the other hand, does not contain clear descriptions of expected performance at different levels. Unless respondents are shown other performances they were being compared against, it would be difficult to convey the information on how the performances were evaluated. As a result, the pedagogical value of the comparative judgment method can be very limited.

In order to resolve all three issues, a new comparative judgment method format, authentic rubric, was proposed. The authentic rubric combines the characteristics of scoring rubrics and the comparative judgment method to overcome the shortcomings of the comparative judgment method and extend the capability of scoring rubrics. The current study aimed to evaluate if the proposed scoring procedure can generate scores that have psychometric properties that were comparable, if not better, to scores collected from scoring rubrics.

1.3 STUDY OVERVIEW

This section provides an overview of this dissertation. The purpose of this study was to examine the psychometric properties of the evaluation procedure that combines the holistic rubric and the comparative judgment method. The new evaluation procedure can theoretically resolve some of the shortcomings inherent in comparative judgment methods while retaining its strength. In order to evaluate this new procedure, an essay set published online by the Hewlett Foundation on Kaggle.com was used. This essay set includes two professional rater scores for each essay, a holistic rubric used by the professional raters, the original prompt, and 1785 student essays. The professional rater scores were used to guide the effort of selecting anchors in the authentic rubric. Two authentic rubrics were created to examine the relation between characteristics of the anchors and the psychometric properties of the rater judgments.

This study recruited participants from the Amazon Mechanical Turk (Mturk) platform to take part in a training session, where the participants were asked to use either a holistic rubric or one of the two authentic rubrics to evaluate a small number of essays. Participants who were able to meet the minimal performance standard during the training session were invited back to evaluate a large subset of the essays in a number of evaluation sessions. Throughout the evaluation sessions, the time taken to evaluate the essays was recorded and raters were asked to report their perception of the ease of use and enjoyment of using the evaluation tools. After the evaluation sessions, both the professional rater scores and the participant raters scores were fitted to the many-facet Rasch model (MFRM; Linacre, 1989, 1993) to obtain the latent trait estimate for each essays. The psychometric properties of the authentic rubric were then examined via inter-rater agreement, agreement with professional rater scores, raters' self-reported ease of use and enjoyment, and time taken to evaluate each essay.

1.4 OVERVIEW OF THE ORGANIZATION

The purpose of this dissertation was two-folded: First, it described a performance assessment scoring procedure that combines the characteristics of comparative judgment and scoring rubrics. The proposed method was designed to overcome the shortcomings of currently available comparative judgment methods and extends the capability of the current best practice in scoring performance assessment. Second, a study was carried out to validate the proposed scoring procedure. The validation effort was guided by the argument-based approach (Kane, 1992, 2013; Kane et al., 1999), where the concern of the validity study was to collect evidence to support the proposed interpretation or use of the scores.

In chapter 2, current state of the art performance assessment procedures in education are reviewed. The chapter describes the procedure and rationale in developing and using holistic rubric in performance assessment. Afterwards, the theoretical foundation, justifications, and empirical support of comparative judgment are reviewed. With comparative judgment method receiving relatively little attention in the educational assessment literature, works of comparative judgment implementation in other disciplines are also reviewed to garner additional empirical evidence. The chapter concludes with a discussion of remaining unresolved shortcomings of the comparative judgment method in educational assessment, and a proposal to address these concerns.

In chapter 3, the argument-based approach to validation is briefly described. The framework proposed by Kane (1992, 2013) is then used to develop arguments in support of the proposed scoring procedure. Hypotheses in relation to the arguments are then proposed. Afterwards, description of the participant recruitment procedure, instruments used, and the essays being evaluated is provided. The chapter ends with a description of the analytic procedures that were used to analyze the collected data. Chapter 4 presents the demographic information of the participants and results of testing the hypotheses using response data from the 22 participants who passed the training session. The reliability, validity, and efficient of the rubric scores along with the rater experience were examined. Finally, chapter 5 discusses the overall findings in relation to the supporting argument for comparative judgment and the implications for performance assessment in the future. Limitations of the study and future area of research are also highlighted.

Chapter 2

2.1 PERFORMANCE ASSESSMENT

As Kane et al. (1999) noted, the term “performance assessment” can be misleading since every assessment would require the examinees to perform some actions. What distinguishes performance assessment from other forms of assessments (e.g., objective assessment) is its open-ended format. By expanding the possible response options, it is believed that performance assessment can be more authentic to real-world problem solving and more reflective of student’s understanding (Bond, 1995). It was also claimed that performance assessment can more easily assess higher-order cognitive skills that objective assessments, such as multiple-choice, have difficulty measuring (e.g., Frederiksen, 1984). However, performance assessment cannot be scored by machines easily and can introduce significant cost. An increasing body of research is devoted to developing algorithms that can utilize computers to assess open-ended responses, such as the quality of written essays, automatically (e.g., Attali, 2015; Kakkonen, Myller, Timonen, & Sutinen, 2005; LaBerge & Samuels, 1974; Larkey, 1998; Rawson & Middleton, 2009). Yet, the work in this area is still in its infancy and there is an argument that the automated approach to essay scoring cannot assess higher-order aspects of writing, such as quality of ideas and organization of the writings (Attali, Lewis, & Steier, 2012). Therefore, human judgment is still invaluable and necessary in providing a holistic view of the performances.

2.2 SCORING PROCEDURES

Early research on performance assessment found that any given essay can receive any score in the range when raters were not given explicit guidance (White, 1985). This lack of consistency in rater judgments was troubling because it raised the question of the generalizability of the scores. In response to the need to raise consistency in rater judgments, various scoring guides with the goal of standardizing the scoring procedure were developed. These methods included scoring rubrics (Johnson et al., 2009), checklists (Johnson et al., 2009), feature analysis (Cooper, 1977), primary trait scoring (Lloyd-Jones, 1977), and general impression marking (Cooper, 1977).

Among the available scoring guides, scoring rubrics, which operationalize the different possible performance levels in the particular domain (Johnson et al., 2009), is probably one of the most widely adopted method. Scoring rubrics can be further differentiated into two approaches, holistic and analytic scoring (Livingston, 2009). In analytic scoring, the raters use a rubric that specifies the dimensions of the assessments and the possible number of points awarded to each dimension (Odell & Cooper, 1980). This scoring format is considered to be particularly useful in classroom settings since it contains diagnostic information about the student's strengths and weaknesses (Jonsson & Svingby, 2007). Holistic scoring, on the other hand, requires the raters to judge the quality of the performance by the overall impression. Thus, a holistic scoring rubric only provides a general description of expected performance at each score point.

There was a debate over the relative merits of analytic and holistic scorings. On one hand, analytic scoring was found to produce higher inter-rater agreement in some cases (Alharby, 2006; Klein et al., 1998), and considered to be more useful in communicating the strengths and weaknesses of each performance to students in classroom (Jonsson & Svingby, 2007). On the other hand, analytic scoring was more costly to implement (Mullis, 1984). It took more resources to train and score the performances using analytic scoring. More importantly, advocates of holistic scoring suggested that the overall evaluation is not the sum of its parts (Lloyd-Jones, 1977; Mullis, 1984; White, 1984, 1985). Focusing on analyzing the features of each performance could lead to the raters overlooking the overall quality of the performance. A more thorough discussion over the merits of the two scoring methods is beyond the scope of the current review. It is important, however, to note that the debate over advantages of the two scoring methods may be purely academic in practice, since research found little difference between the two scoring methods in psychometric properties (Swartz et al., 1999), especially if the analytic scores were aggregated (Alharby, 2006; Goulden, 1994; Klein et al., 1998). In the current review, development of holistic scoring is discussed in further details since the proposed authentic rubric shares a similar format with holistic scoring rubric.

2.2.1 Holistic Rubric

Holistic scoring was developed by the Educational Testing Service (ETS) in response to the need of a scoring method that can result in consistent essay scorings (White, 1985). A holistic scoring rubric is typically consisting of a guide that specifies the range of scores that could be assigned to the performances. Each score level is also accompanied by a description of the features to be evaluated and the expected quality of each feature. Yet, while the scoring rubric is designed with the intention of providing detailed descriptions to aid the raters in differentiating levels of performance, the complex nature of the task limits the descriptions to be abstract at best (Lumley, 2002). Therefore, a crucial procedure in constructing holistic scoring rubric is the rangefinding step (Johnson et al., 2009). During rangefinding, example responses to the performance tasks are collected to identify the range of possible performances. The example responses will then be evaluated by a panel of experts to ensure the holistic scoring rubric accurately describes the range of real world performances. Furthermore, a subset of the example responses would then be chosen to serve as benchmarks, which are also referred to as anchors (Mullis, 1984). Benchmarks may be chosen to represent the range of responses across the scale, the typical or average performance at each score point (Johnson et al., 2009; Wolcott & Legg, 1998), or the borderline cases (e.g., responses that barely qualified for a score or just missing a score; Livingston, 2009). Benchmarks are important rater training materials in that they serve as the exemplar of each score level and can be used to monitor rating accuracy in evaluation sessions.

2.3 COMPARATIVE JUDGMENT

The "Law of Comparative Judgment" developed by Thurstone (1927a) lays the theoretical foundation of all judgment processes that involve comparisons of two or more objects. In his original research, Thurstone was interested in developing scaling methods in psychophysics that can measure human judgments or perceptions of physical stimuli, such as weight, length, and size of various objects. Thurstone (1959) later generalized the methodology to develop scales that measure latent construct such as attitudes, values, or tastes.

According to Thurstone (1927a, 1927b), when a stimulus is presented to a judge, the stimulus will elicit a psychological response. The psychological response is named *discriminal process*. It is further assumed that, all things being equal, a different discriminal process is produced every time the judge is presented with the same stimulus. The discriminal process can therefore be conceptualized as a distribution, where the most common process is called the *modal discriminal process* and the standard deviation of the distribution is named *discriminal dispersion*. The perceived difference between two stimuli is assumed to be a function of the difference in the corresponding modal discriminal processes, named *discriminal difference*. Finally, Thurstone developed the measurement model to estimate the discriminal difference between objects by tallying the pairwise comparison decisions provided by the participants.

Since Thurstone (1927a) introduced his law of comparative judgment, the methodology was adopted by psychology related disciplines including measurement, behavioral economics, personality psychology, and industrial-organization (I-O) psychology. In the literature, the comparative judgment method has been referred to as ordinal comparison (e.g., Shah et al., 2013), pairwise comparison, and relative judgment (e.g., Goffin & Olson, 2011). The implementation of the method also varies across disciplines and studies where some adopts the original pairwise comparison design (e.g., Heldsinger & Humphry, 2010; Jonesa & Alcocka, 2013; Pollitt, 2012), others used the rank order design (Attali, 2014), and some implemented a design similar to a Likert-type item (e.g., Goffin et al., 1996; Wagner & Goffin, 1997). In general, when the pairwise comparison design was implemented, participants would be asked to compare a set of items against each other to determine which items better satisfy a specified criterion. For instance, Thurstone's (1927a) early work in the psychophysics domain would ask participants to compare a series of lines and determine which line is longer. In the attitude domain, participants would be asked to compare a set of statements and determine which statement reflects a more favorable attitude towards the subject. With a sufficient number of pairwise comparison decisions, the items or objects can be rank-ordered. The scaling procedure developed by Thurstone can then be used to place the items on a common linear scale. Later, Andrich (1978) noted the similarities between Thurstone's measurement model to Rasch measurement theory and provided the mathematical proof of the equivalency.

2.3.1 Comparative Judgment in Education Assessment

There are essentially two lines of research in educational assessment that investigated the reliability and validity of the comparative judgment method in performance assessment. The first line of research was originated in England, where Pollitt and Murray were credited for conducting the first study implementing comparative judgment in spoken language assessment (Pollitt, 2004). Afterwards, the comparative judgment method was used to equate test scores collected in different years (Bramley, Bell, & Pollitt, 1998). Pollitt (2004) later outlined a comprehensive proposal of using comparative judgment to evaluate writings. The advocates of the comparative judgment method argued that it could, by design, eliminate systematic rater biases (Bramley, 2007). Furthermore, it was believed that the raters would be relieved from the burden of scaling their judgments since raters would no longer be responsible for determining the distance between each score point on the scoring rubrics (Attali, 2014; Bramley, 2007). The research in England had subsequently inspired an increasing body of research to examine the feasibility and merits of comparative judgments method in educational performance assessment (e.g., Attali, 2014; Jonesa & Alcocka, 2013; Jones & Inglis, 2015; Jones et al., 2014; Pollitt, 2012; Seery, Canty, & Phelan, 2012).

Another line of research in comparative judgment method was inspired by the increasing popularity in massive open online courses (MOOCs). Instructors found that hundreds, if not thousands, of students would enroll in MOOCs due to the low barrier of entry. Among many proposed methods to handle the large volume of student submitted assignments, one proposed solution was to rely on peer assessment in lieu of instructor grading. Although peer assessment can potentially resolve the logistic issue in MOOCs, some scholars raised concerns over the reliability and validity of peer assessment. In particular, peer assessment in MOOCs often limited the number of evaluations conducted by each peer rater to be fairly small. Estimating and correcting for systematic rater biases can therefore be difficult. In this context, comparative judgment, or ordinal comparison, was proposed as a solution that can elicit more accurate peer judgment, and eliminate systematic biases by the scoring procedure (Shah et al., 2013). A number of studies have since been devoted to examining the validity of the method and the statistical models that are most suitable for modeling ordinal data (Caragiannis, Krimpas, & Voudouris, 2014; Mi & Yeung, 2015; Raman & Joachims, 2014, 2015; Sajjadi, Alamgir, & von Luxburg, 2015; Shah et al., 2014). It is, however, important to note that not all research conducted in this area is considered to be relevant for the current study. Some studies are considered irrelevant in the current context since they did not collect rater judgments through comparative judgment tasks (e.g., Mi & Yeung, 2015; Sajjadi et al., 2015; Shah et al., 2014). Instead, they inferred the comparative judgment decisions by using rater judgments collected through the use of scoring rubrics. For instance, if raters assigned a score of 5 and 3 for two essays, it was assumed that raters would judge the essay received a score of 5 to be superior than the essay received a score of 3 using the comparative judgment method. This data conversion procedure means that the converted comparison data would always contain less information than rating data by converting the data from interval scale to ordinal scale.

2.3.2 Comparative Judgment in Other Fields

As previously mentioned, in addition to being adopted in educational assessment, the comparative judgment method was also used in a variety of contexts. For instance, the comparative judgment method was used to study the rankings of risk perceptions (Florig, Morgan, Morgan, & Jenni, 2001; Morgan, DeKay, Fischbeck, & Morgan, 2001), food characteristics (Oakes & Slotterback, 2002), clinical services (Hazell, Tarren Sweeney, Vimpani, Keatinge, & Callan, 2002), lecturer qualities (Wagner & Goffin, 1997), job performances (Goffin et al., 1996, 2009), and attitudes (Olson et al., 2007).

There is one specific line of research that is worthwhile to mention due to its unique implementation of the comparative judgment method. In most comparative judgment studies, raters were asked to compare two or more items explicitly. In contrast, Goffin and colleagues (Goffin et al., 1996, 2009; Olson et al., 2007; Sheppard, Goffin, Lewis, & Olson, 2011; Wagner & Goffin, 1997) elicited comparative judgment with a format similar to Likert-type items. Specifically, they used the Relative Percentile Method (RPM), which utilized an item with a 101-point scale. Although the RPM shared the same format as Likert type item, it was argued to be unique in two aspects. First, interpretation of the scale of the RPM was different from a Likert-type item, where each point on the RPM represented a percentile score. For example, if the rater assigned the ratee a 70 point on the scale, it implied that the rater believed the ratee was better than 70 percent of other people being evaluated. Therefore, the RPM scale arguably elicited relative judgment instead of absolute judgment. The second unique characteristic of the RPM was that all ratees or performances were judged on one item. The raters were explicitly asked to evaluate all ratees on the same item. As a result, the raters would be cognizant of the relative position of all ratees on the scale.

The RPM implementation of comparative judgment was significantly different from the pairwise comparison design. It avoided the laborious process to compare every other performance, and criterion validity studies in I-O psychology, attitude research, and perception of others found promising results (Goffin et al., 1996, 2009; Olson et al., 2007). Thus, this specific methodology could potentially circumvent one of the shortcomings of comparative judgment method. This issue that will be further discussed in later sections.

2.3.3 Theoretical Advantages

The arguments supporting the use of comparative judgment in performance can largely be summarized as follows: First, it was believed that comparisons are inherent in all social judgments (Mussweiler, 2003). In fact, it was argued that the perceived level of performance is registered as stronger or weaker than others instead of being comprehended in absolute terms (Goffin & Olson, 2011). Therefore, the comparative judgment method is more aligned with cognitive processes involved in evaluations, which can then potentially lead to more accurate assessments.

Second, it was argued that the comparative judgment method can establish a common frame of reference among the raters (Böckenholt, 2004; Goffin & Olson, 2011). When scoring rubrics or Likert-type items are used, it is unclear if the raters would share the same interpretations for all the response categories. The interpretation of what *excellent* or *strongly agree* represents may differ across raters. In a comparative judgment task, the frame of reference is represented by other performances. Thus, it is expected that the arbitrary nature and idiosyncratic interpretation of rating scales will be removed.

Third, the comparative judgment method was argued to relieve the raters from the burden of scaling their judgments (Attali, 2014; Bramley, 2007). When raters are using the scoring rubrics, the raters have to decide the difference between each score points. In order to determine which score a ratee should receive, the raters will have to first memorize or internalize the difference between the score points. Then, the raters will have to determine which score point better matches the performance. In contrast, the comparative judgment method would not require raters to internalize the scale of measurement. They are only required to determine the relative quality of the performances.

Fourth, the comparative judgment method can, by design, eliminate all systematic biases, such as rater severity and item difficulties, in performance assessments (Bramley, 2007; Linacre, 1989; Shah et al., 2013). Regardless of the severity of the raters, the rank ordering of the performances obtained through the comparison method should remain the same. This property of comparative judgment can be illustrated with the formulation often used in modeling choice behavior as follows:

$$D_1 = T_1 + R_j \tag{1}$$

where D_1 represents the discriminal process elicited by performance 1, T_1 represents the true score of performance 1, and R_j represents the systematic bias of rater j. Same formulation can be applied to performance 2. In the case of comparative judgment, each decision is determined by the discriminal differences, which can be represented as the difference between the two discriminant processes,

$$D_D = D_1 - D_2 \tag{2}$$

$$D_D = T_1 + R_j - (T_2 + \ R_j) \tag{3}$$

$$D_D = T_1 - T_2 \tag{4}$$

as shown in equation 4, the systematic rater bias is eliminated in the comparative judgment. This characteristic of comparative judgment is argued to be particular important in peer ratings in MOOCs where each peer raters only evaluate a handful of the responses (Shah et al., 2013).

2.3.4 Empirical Support

If comparative judgment holds many theoretical advantages over scoring rubrics, it is expected that some of the theoretical advantages will translate into rater judgments that are more consistent and accurate. In fact, studies comparing the psychometric properties of comparative judgment method and scoring rubric mostly reported results favoring the comparative judgment method. Studies in other disciplines found the use of comparative judgment method produced evaluations that were less prone to random error (Shah et al., 2014, 2013) and had higher criterion validity (Goffin et al., 1996, 2009; Olson et al., 2007; Sheppard et al., 2011; Wagner & Goffin, 1997). Educational research findings, on the other hand, were mixed. The current section attempts to summarize currently available empirical evidence and point out the plausible cause for mixed findings reported in educational research.

The series of exploratory studies conducted by Shah and colleagues (Shah et al., 2014, 2013) provided some unique data in examining the accuracy of comparative judgment. In these studies, participants were asked to either compare the relative quality of two objects or determine the absolute quality of the two objects. As an example, the participants would be presented with two circles of different sizes and asked to judge which circle is larger (the comparative judgment condition), or to give an estimate for the size of each circle (the absolute judgment condition). In the absolute judgment condition, estimates provided by the participants would then be converted to comparison decisions. For instance, if the participants claimed that the size of the first circle is 10 whereas the size of the second circle is 5, the record would show that the participants determined the first circle to be larger than the second circle. The absolute judgments and comparison judgments derived from it were then evaluated against an accepted gold standard (e.g., the area of each circle). Shah and colleagues found the comparison judgment condition yielded evaluations that had either the same or higher degree of accuracy among seven rating tasks used in the studies.

Similar findings supporting the accuracy of comparative judgments were also reported in studies involving job performance evaluations. Wagner and Goffin (1997) recruited 80 undergraduate students (i.e., novices) and 14 graduate students (i.e., experts) to evaluate the speaking style and clarity of four lecturers with either the RPM or Likert-type items. The raters were asked to evaluate the two qualities of lecturers holistically or analytically. Using the accuracy analysis framework proposed by Cronbach (1955) and average expert evaluations as the gold standard, Wagner and Goffin found that novice raters who used the RPM provided evaluations that were more accurate in discriminating the lecturers within each criteria. Similar findings were reported in a related study where Sheppard, Goffin, Lewis, and Olson (2011) investigated the accuracy of the comparative judgment method in evaluating job interview performances. Using the same research design that involved novice and expert raters, Sheppard and colleagues found that the evaluations collected through RPM were significantly more accurate at discriminating the performances on criteria being assessed.

It is important to note that the empirical support for the validity of comparative judgment method was not limited to the higher degree of correspondence between novice and expert raters. Goffin et al. (1996) examined the criterion-related evidence of validity of comparative judgment methods and rubric scorings by collecting job performance assessments of 88 employees from 27 district managers of a company. Additionally, they collected survey data and cognitive test results from the employees. Examining the correlations between the managers' assessments and different criteria, they found the comparative judgment scores, collected through RPM, had higher correlations with criteria such as personality and vocational interest, verbal and quantitative cognitive abilities.

Finally, Goffin et al. (2009) examined if the comparative judgment method yielded more information than scoring rubrics by comparing the predictive power of RPM and scoring rubric in job performance assessment. One hundred and seventy (170) participants in the study were assessed on a set of job-related skills using scoring rubrics. After a year, the supervisors of the participants were asked to evaluate each participant's job performance using either the RPM or scoring rubrics. Then, Goffin and colleagues examined the explanatory power of the two assessment methods using a series of hierarchical regressions. The regression models showed that both scoring procedures produced evaluations that were significant predictors of job performance. However, Goffin and colleagues found the comparative judgments predicted variances above and beyond the scores produced by the use of scoring rubrics. With rubric scores entered as predictor, adding the scores obtained through comparative judgment into the regression model improved the predictive power of the model significantly. Reversing the order of the predictors entered into the model revealed that scores from scoring rubrics did not improve the explanatory power of the model above and beyond comparative judgment scores. Goffin and colleagues therefore concluded that comparative judgment method provided more pertinent information than scoring rubrics.

In the context of educational assessment, researchers examined the use of comparative judgment in evaluating mathematics problem solving (Jones & Inglis, 2015), chemistry knowledge (McMahon & Jones, 2014), peer assessment (Jonesa & Alcocka, 2013), narrative writing (Heldsinger & Humphry, 2010), and short writing tasks (Attali, 2014). As previously mentioned, although the studies reviewed thus far generally found the comparative judgment method yielding stronger evidence of validity, the findings in educational assessment were mixed. Comparing the correlations between different performance assessment methods and external criteria, such as prior achievements, or expert ratings, studies in education reported the estimations computed through comparative judgment had higher (McMahon & Jones, 2014), similar (Attali, 2014; Jones & Inglis, 2015), or lower (Attali, 2014) correlations than using scoring rubrics. One of the potential explanations for the inconsistent findings in educational assessment could be that the studies used different comparative judgment methods and measurement models. In the studies where comparative judgment reported to have higher correlations with external criteria, the authors used the pairwise comparison design and Rasch model to estimate the scores of each ratee (Jones & Inglis, 2015; McMahon & Jones, 2014).

In the study conducted by McMahon and Jones (2014), students responded to questions examining their knowledge in carrying out chemistry experiments. Teachers were instructed to use either scoring rubric or pairwise comparison to evaluate student responses. The comparative judgments were then modeled by the Rasch model. Additionally, McMahon and Jones collected the students' performance on four chemistry tests administered in the past. Comparing the correlations between achievement tests and evaluations collected through comparative judgment and scoring rubric, it was found that the comparative judgments had a higher correlation with prior achievements (r = .536) than scoring rubric (r = .351).

Jones and Inglis (2015) examined the use of comparative judgment in assessing mathematical problem solving skills. Using a design similar to the study conducted by McMahon and Jones (2014), Jones and Inglis collected expert evaluations through the use of pairwise comparison and scoring rubric. In the comparative judgment condition, 15 professional raters were assigned to complete 250 to 300 judgments, whereas 2 raters in the rating condition evaluated all the students' works using a 50-point holistic rubric. Correlational analysis revealed that the two types of judgments had similar correlations with student's predicted standardized test grades.

In contrast to the two previously reviewed studies, Attali (2014) implemented a comparative judgment design where participants, instead of engaging in pairwise comparisons, ranked a subset of writing responses all at once. The participants were required to rank multiple sets until all responses had been ranked at least once and it was able to infer an overall rank of all the responses. A model modified from the Elo rating system (Elo, 1978) was then fitted to the rank orders to obtain an estimate of the quality of the responses. Depending on the parameter values used and amount of prior information incorporated in the model, Attali found the estimations from rank orders had either similar or lower correlation with expert ratings than scores obtained from holistic rubric.

In summary, the exploratory studies conducted by Shah and colleagues (Shah et al., 2014, 2013) investigated the accuracy of comparative judgment with a number of artificial tasks. By using tasks that had objective measurements readily available, Shah and colleagues were able to examine the evaluation accuracy of the two types of judgments, and their study findings suggested that evaluation accuracy of comparative judgment was either similar or higher than absolute judgment. The findings reported by Shah and colleagues were consistent with a series of studies conducted by Goffin and colleagues (Goffin et al., 1996, 2009; Sheppard et al., 2011; Wagner & Goffin, 1997) that compared the criterion-related validity evidence of comparative judgment and scoring rubrics. The studies used RPM showed a clear pattern that comparative judgments had a higher agreement with expert judgments, higher correlations with external criteria, and higher predictive power than scores obtained through scoring rubrics. Finally, although studies in educational assessment had not examined the issue extensively, the results of existing studies were generally consistent with the findings reported in the other disciplines. Specifically, it is found that the use of comparative judgment to evaluate written responses had a higher correlation with prior achievement than using holistic rubric.

2.3.5 Remaining Issues

The literature reveals that comparative judgment can yield comparable, if not stronger, evidence of validity than using scoring rubrics. However, a successful implementation of the comparative judgment method in educational assessments will need to resolve several practical issues. First, when comparative judgments are implemented in the pairwise comparison design, the raters are required to carry a workload that is significantly higher than using scoring rubrics.

For instance, McMahon and Jones (2014) noted that in order to evaluate 154 student reports, the original pairwise comparison design would require the raters to make 11,781 comparison decisions. In an attempt to lessen the workload, McMahon and Jones reduced the number of comparison decisions to 310 per raters, since it is found that a relatively stable estimations can be obtained with small samples of rater judgments (Pollitt, 2012). Nonetheless, 310 comparison decisions were still ten times the workload if the raters were simply asked to score the 31 reports.

Different solutions were offered to reduce the number of comparison decisions. Pollitt (2012) developed an adaptive system with a structure similar to computerized adaptive testing. The adaptive comparative judgment system would estimate the ability of ratees after each comparison decision was submitted by the raters. Based on the estimation precisions, the algorithm would determine which ratee requires additional judgments. By adjusting the threshold for acceptable estimation precisions, the adaptive comparative judgment system could reduce the number of judgments required in comparative judgment substantially. Another approach, proposed by Attali (2014), was to ask the raters to rank order a set of performances instead of making separate comparison decisions for every possible pairings. Pairwise comparison decisions could then be extracted from the rank orders by assuming that the pairwise comparison decisions would be consistent with the rankings.

Although the aforementioned proposals can potentially resolve the first shortcoming of the comparative judgment method, they failed to address the second shortcoming, which is that pairwise comparison designs do not convey information on the absolute quality of the performances being evaluated (Böckenholt, 2004). The rank orders of performances only communicate the information on which performance is considered to be the best or worst; however, it does not contain information on whether any of the performances is actually of high or low quality. It could be the case that the best performance in the set still failed to reach the minimum level of competence. In order to determine the absolute quality among the ranked responses, it was proposed that ratings and comparative judgment can be combined (Böckenholt, 2004). For example, raters may be asked to use a Likert-type item to rate some of the ranked responses. Using the ratings as the anchor points, it would be possible to infer the absolute quality of the ranked responses. Alternatively, raters can, after determining the rank orders of the performances being evaluated, retroactively examine which performance reaches the expected level of performances.

Nonetheless, none of the proposed solutions to the first and second concerns resolve the third problem associated with pairwise comparison design. Specifically, the comparative judgment method cannot effectively inform the ratees on the evaluation standards. Using scoring rubrics can inform students and teachers on the evaluation standards and increase transparency in the assessment process. Further, teachers can use scoring rubrics as a pedagogical tool by pointing out how students can improve to reach certain performance level. In comparative judgment, unless all the performances under evaluations are released to the students, it is difficult, if not impossible, to inform ratees the reasoning behind the scores assigned to them. More importantly, ratees will not have the opportunity to learn from their mistakes and improve their performances since they will not able to examine how they reach, or fail to reach, the expected performances.

2.4 PROPOSED SOLUTION: AUTHENTIC RUBRIC

To create a solution that can address all the shortcomings of comparative judgments, it is proposed that scoring rubrics can be modified to incorporate comparative judgments, thereby retaining the strengths and overcoming the weaknesses of the comparative judgment method.

This newly proposed comparative judgment method is best termed authentic rubric since it retains the scoring rubric format while emphasizes the role of example performances. Specifically, it is proposed to use benchmark performances to represent each performance level in a scoring rubric and abandon the need to label each performance level with numerical value. The raters, during training and evaluation, do not have to be aware of the numerical values associated with each performance nor the length of the rubric. Instead of determining which score point best matches the performance being evaluated, the raters will be asked to compare the relatively quality of the benchmark performances against the ratees' performances. Measurement models, such as the MFRM, can then be fitted to the comparative judgments to obtain estimations of ratees' abilities.

Using writing assessment as an example, it is necessary to develop a scoring rubric that describes the expected level of performance at each score level. In addition to providing detailed descriptions at each score level, development of a scoring rubric for writing assessment is often accompanied by a rangefinding procedure where the goal is to collect example responses to the prompt (Johnson et al., 2009). The experts will then evaluate the example responses to determine if the scoring rubric accurately describes the range of responses and which example responses can serve as benchmarks at each score level. The benchmarks and scoring rubric are then used in training the raters to ensure raters can accurately discriminate between performance levels (Johnson et al., 2009; Mullis, 1984; Osborn Popp, Ryan, & Thompson, 2009; White, 1984).

Development of the authentic rubric will follow the same procedure where benchmark papers are located. The difference between the proposed authentic rubric and a conventional scoring rubric is how these benchmarks are utilized. For conventional scoring rubrics, the benchmarks are essentially training materials to supplement the rubric. Conversely, the benchmarks in the authentic rubric method embodies the expected performance levels. They are not only training materials for raters; rather, the benchmarks are directly used in the evaluation process. In the scoring session, the raters will be asked to compare the relative quality of benchmarks and respondents' writings. The comparative decisions, in turn, are modeled by the statistical measurement model.

2.4.1 Advantage Over Pairwise Comparisons

Unlike currently used comparative judgment methods, the authentic rubric addresses all three shortcomings. First, the authentic rubric method does not involve a large number of pairwise comparison decisions. The reason for the number of comparison decisions to grow exponentially in the pairwise comparison design is that the number of comparison required increases geometrically as the number of responses grows arithmetically. The authentic rubric design circumvents the problem by predetermining the number of expected performance levels and the corresponding number of comparisons. In this sense, the format of authentic rubric is similar to that of RPM where the length of the scale is determined by ad hoc decisions (Wagner & Goffin, 1997).

Second, the benchmark performances in authentic rubric can be chosen to represent absolute performance levels. Therefore, each comparison decision not only entails the relative quality of the performance; it also contains information on whether the performance has exceeded certain performance threshold. If a performance is judged to be better than the benchmark that resembles excellent performances, the performance can then declared as having superior quality. Although a post hoc decision in pairwise comparison designs can achieve the same goal, the advantage of making an ad hoc decision lends itself to the third advantage.

Lastly, with the benchmark performances chosen prior to the evaluation process, they can be used to demonstrate the expected performances to test users. It clarifies the expectations for the assessment, which is valued by both teachers and students (Schamber & Mahoney, 2006). The raters can point out which benchmark performances are considered to be superior and inferior to the respondents' performances. Subsequently, respondents can study the benchmark performances and determine the improvements they may able to make.

2.4.2 Advantage Over Scoring Rubrics

There are three arguments supporting the use of a rubric that illustrates performance levels with benchmarks and solicit evaluations through direct comparisons. First, if we assume the various theoretical advantages of comparative judgment method are the underlying cause of the reported stronger evidence of validity in previous studies, then the proposed authentic rubric should be able to leverage some of the strengths of comparative judgment and produce evaluations that are more accurate than using scoring rubric. Further, the authentic rubric, sharing similar format as scoring rubrics, should able to collect evaluations that are at least comparable to the ones collected using scoring rubrics.

Second, authentic rubric places a heavier emphasis on the role of the benchmark performances. Training raters in using scoring rubrics is often accompanied by a set of benchmarks, which are supposed to be example performances representing each score point (Johnson et al., 2009). Studies have found that benchmarks had a significant impact on the scores assigned by the raters (Hughes & Keeling, 1984; Osborn Popp et al., 2009). Different benchmark papers chosen to represent each score point resulted in substantially different item difficulties as estimated by MFRM. Furthermore, the differences in assigned scores would lead to substantially different classification decisions in terms of determining whether the students met the standards set forth by the testing program (Osborn Popp et al., 2009). Although identifying benchmark performances are regarded as one of the important steps in developing scoring rubrics, research in this area was lacking (Osborn Popp et al., 2009). By incorporating the benchmark performances as key components of the evaluation process, the authentic rubric calls attention to the important, yet, often overlooked factor in performance assessment.

Third, authentic rubric affords the flexibility in how rubrics can be implemented. Training raters to use scoring rubrics requires the raters to internalize all score points in the scale. Typically, scoring rubrics use 4- to 6-point scales to avoid confusions across score levels. The authentic rubric, on the other hand, decouples the numeric scores from the rubric. As a result, longer scale can be segmented where each rater is only responsible for a subset of the rubric. For instance, assuming a rubric has a 10-point scale. Traditional training will require the raters to internalize all ten points on the scale and successfully discriminate them. It makes little sense to train and ask raters to only use a subset of the scale. However, for authentic rubric, raters do not have to be aware of the length of the scale. The 10-point scale can be represented by two sets of anchors with raters being trained to use one of the them.

Chapter 3

3.1 METHODS

Using previously developed holistic rubric and collected expert ratings, this study developed three scoring rubrics. Amazon Mechanical Turk (Mturk) was used to recruit participants to evaluate essays using one of the three methods. The evaluation decisions provided by the participants were fitted to the MFRM to obtain latent trait estimates of the essays being evaluated. These estimates would then be evaluated by examining its accuracy in reproducing the gold-standard; namely, scores provided by professional raters. The reliability of rater judgments and experience of using the evaluation tools were also examined.

An important purpose of this study was collecting validity evidence to validate the authentic rubric developed in the study. The validation framework proposed by Kane (1992, 2001, 2004, 2013) was used to guide the validation effort. Kane's framework emphasized that validation effort should be focused on collecting evidence to support the proposed use and interpretation of the scores called interpretive argument. The interpretative arguments play the role of a formal theory by clearly specifying the underlying assumptions of the interpretations. Therefore, the inferences and assumptions of the interpretative arguments need to be supported in order to claim that the proposed interpretation and use of score are valid.

There are several components of argument-based validation that are common in test-score interpretations including scoring, evaluation, generalization, extrapolation, and decision making (Kane, 2004). Three of these components: scoring, generalization, and extrapolation, are particularly pertinent to this study due to the fact that collecting validity evidence to support the scoring procedure is essential to establish these components of validity argument. The scoring component requires evidence to support the assessment, data collection, and scoring procedure design. The generalization component involves evidence that can support generalizing the observed score to a trait value. The extrapolation component entails collecting evidence to link the data collected in the test to the behaviors of interest in the real world. The inferences, assumptions, and evidence needed to establish these three components of validity arguments and pertinent to scoring procedure are listed in table 3.1.

Table 3.1. Framework for Validity Argument – Inferences, Assumptions and Evidence Needed

Inferences	Assumptions	Evidence needed
Inference 1 Scoring: from the observed performance to the observed score		
Assumption 1.1	The scoring criteria are appropriate and acceptable	The anchors selected for the authentic rubric will be representative of the scoring categories.
		Criterion validity evidence (i.e., correlations with gold standard scores) will be collected to ensure the anchors provide meaningful discrimination among performance levels.
Assumption 1.2	Raters will be able to score the responses reliably	The raters will be trained to use the authentic rubric using a previously verified procedure (Attali, 2015). Raters will be expected to achieve minimum level of consensus after completing the training session.
Inference 2 Generalization: from the observed score to the expected universe score		
Assumption 2.1	The scores given by the raters will be consistent, and random errors due to raters will be controlled.	The scoring procedure will be standardized to control for random error caused by administration occasions.
		Multiple raters will be used to assess each performance to reduce the impact of personal biases.
		Comparative judgment method was argued to be capable of eliminating systematic rater biases by design. MFRM will be fitted to the evaluations to determine if the claim

		is warranted and the potential impact of rater biases. The inter-rater agreement will be examined to ensure the ratings are consistent across raters.
Inference 3 Extrapolation: from the universe score to the expected level of skill in the target domain		
Assumption 3.1	There are no systematic errors that are likely to undermine the extrapolation. There are either no or equal rater drafts across evaluation sessions.	The MFRM estimates and controls for rater biases, a form of systematic errors.

3.1.1 Hypotheses

Five hypotheses are proposed for the current study. First, it is expected that the evaluations elicited through the authentic rubric procedure will have higher reliability than using traditional rubrics. While holistic rubric scoring categories are subject to individual interpretations, it was argued that comparative judgment method can establish a common frame of reference among the raters (Böckenholt, 2004; Goffin & Olson, 2011). Therefore, it is expected that raters who use comparative judgment methods would have a higher degree of agreement in their interpretation and use of scoring categories.

Second, it is hypothesized that, in comparison to the traditional rubric, using the authentic rubric in evaluation will result in higher validity. Specifically, the authentic rubric evaluation will have a higher correlation with the gold-standard scores provided by the professional raters than traditional rubric evaluation. There are several studies (McMahon & Jones, 2014; Sheppard et al., 2011; Wagner & Goffin, 1997) that found empirical evidence in support of the superior validity of comparative judgment scores. Examination of accuracy of comparative judgment in evaluating physical dimensions also found promising results (Shah et al., 2014).

Third, the use of authentic rubric is expected to improve the rater experience and improve the efficiency of the evaluation process. Thus, it is hypothesized that raters are expected to find using comparison rubrics to be easier and more enjoyable than using traditional rubric. Further, the time taken to evaluate each essay is expected to be shorter for authentic rubric. Advocates of comparative judgment methods argued that the comparison process is inherent in all evaluations. Additionally, comparative judgment methods are argued to be relieving the burden of internalizing the difference between score points (Attali, 2014; Bramley, 2007). Combination of both factors is expected to ease the mental effort required to evaluate essays and increase the efficiency of the process.

3.1.2 Participants

A total of 55 participants were recruited through the Amazon Mechanical Turk (Mturk) platform. The MTurk is a crowdsourcing marketplace where researchers are allowed to recruit participants with monetary compensation. Study found participants recruited from the Mechanical Turk platform is more representative of the convenient sample recruited from universities (Berinsky, Huber, & Lenz, 2012). Further, participants recruited from the platform in previous studies successfully completed similar evaluation tasks (e.g., Attali, 2014, 2015).

3.1.3 Materials

The study essays used in this study was originally published by the Hewlett Foundation on Kaggle.com. The original dataset consisted of eight essay sets written by students from grade 7 to grade 10. In this study, only one of the essay sets was used. One hundred and twenty-seven

(127) out of the 1785 essays were selected to be used in this study. A detailed description of the selection criteria for the 127 essays is provided below.

Prompt. The essay set asked the students to write a letter to the local newspaper to state their opinion on the effects of computers on people. Students were expected to write a few sentences to respond to the prompt. The original essay prompt can be found in Appendix A.

Holistic Rubric. All the student essays in the set were graded using a 6-point holistic rubric. The rubric is focused on four elements of student essays: amount and specificity of the information, organization of the essay, fluency of the language used, and awareness of the audience. More detailed descriptions of the expected performance for these four elements are accompanied with each score point. The holistic rubric can be found in Appendix B.

Gold Standard Scores. Two professional raters hand-graded the student essays using the 6-point holistic rubric. When the two raters disagreed by 1 point or less, the final score for the essay was calculated as the average of the two scores. However, if the two raters disagreed by more than 1 point, a third rater was asked to give a final score for the essay.

Anchor Essays and Authentic Rubrics. Six (6) essays were chosen randomly to represent each score point of the holistic rubric. The only criterion for the essays to be chosen was that the two raters had to be in perfect agreement in their evaluation. Two authentic rubrics were created using the 6 chosen essays. The first authentic rubric, termed odd scores authentic rubric, was comprised of essays representing score points of 1, 3, and 5; whereas the second authentic rubric, termed even scores authentic rubric, was comprised of essays received a 2, 4, and 6 point in the final score. Two study conditions were created to eliminate the Hawthorne effect of introducing a novel evaluation procedure. Additionally, previous studies found essays received higher scores tended to be longer (Attali, 2014), have more content, better organization, more sophisticated structure, and better writing mechanics (Barkaoui, 2010; Freedman, 1979; Rafoth & Rubin, 1984). With the anchors used in the even scores authentic rubric condition, on average, received higher scores, anchors in the even scores authentic rubric were expected to be longer, contain more argumentative ideas, and have better organization. In other words, it was expected that more essay features would need to be taken into consideration by the raters using the even scores authentic rubric. The anchors used in the two authentic rubrics can be found in Appendix C and D, while the authentic rubric can be found in Appendix E.

Training Essays. Twenty (20) essays with perfect rater agreement were chosen randomly to be used in the training session.

Evaluation Essays. One hundred (100) essays were chosen randomly to be used in the evaluation sessions. These 100 essays were further divided randomly into 5 sets with 20 essays in each set.

Enjoyment, Ease of Use, and Time Taken. Time taken to evaluate each essay in terms of the seconds it took the participants to submit the evaluation was recorded for all participants. Additionally, participants were asked to report the enjoyment and ease of use of the assessment tasks at the end of each evaluation session (see Appendix F).

Demographic Survey. A demographic survey that asked for participants' age, gender, educational level, teaching and grading experience were also administered. The items used in the demographic survey can be find in Appendix G.

3.1.4 Procedure

When participants signed up for the study, they were asked to first complete the demographic survey. After participants completed the survey, they were randomly assigned to one of the three study conditions: rating condition, where participants used the holistic rubric;

authentic rubric one, where participants used the odd scores authentic rubric; authentic rubric two, where even scores authentic rubric was used. With the authentic rubric conditions, the anchors were presented in the order of the final score but the scores received by the anchors were *not* revealed to the participants. Both training and evaluation sessions instruction were modified to fit the evaluation tool used in each study condition.

Training Session. The training procedure used by Attali (2015) was adopted in this study. The participants were first introduced to the essay prompt and the rubric appropriate for their study condition. After the participants had the opportunity to study the essay prompt and the rubric, an example task similar to the training session tasks was shown. In the rating condition, participants were instructed to assign a score to the example essay using the holistic rubric. In the authentic rubric conditions, participants needed to indicate whether they think the example essay is worse than, similar to, or better than the anchors in the authentic rubric. After completing the example task, participants were told that they would evaluate 20 additional essays using the same procedure. Throughout the training session, participants had access to the full rubric or the anchors if they needed it for reference. Additionally, feedback was provided regarding the accuracy of their evaluations. Three (3) points would be awarded if there was an exact match between the assigned score and the gold-standard score and 1 point for a one-point discrepancy. Whenever participants failed to achieve perfect agreement with the gold-standard score, they would be advised to reveal the essays being evaluated and the rubric to improve the accuracy of their evaluations.

At the end of the training session, only participants who were able to obtain a score of 35 or higher during the training session were invited to participate in the evaluation sessions. The score needed to pass the training session, however, was not revealed to the participants. Rather, participants were only notified that they would need to reach minimum performance achievement standard, in the form of high agreement with the gold-standard, to qualify for additional study sessions. The training session, albeit brief, was shown to be effective in increasing rating accuracy for newly trained raters (Attali, 2015) .

Evaluation Sessions. For participants who were able to meet the minimum performance standard, they were invited to participate in 5 additional evaluation sessions. Each evaluation session required the participants to evaluate 20 essays using the same methodology in the training session except that no feedback was given to the participants regarding their evaluation accuracy. At the end of each evaluation session, participants were asked to report the perceived ease of use and enjoyment of the assessment tool. Participants were free to complete the evaluation sessions at their own pace, but all evaluation sessions were completed within 72 hours after the completion of the training session.

Finally, both training and evaluation sessions were completed online via websites developed by the researcher. The participants and researcher had minimal contact throughout the study regarding the scoring procedure.

3.1.5 Analysis

The MFRM was used to model the evaluations provided by the raters. The MFRM is an extension of the Rasch model that accounts for situations which have more than two facets interacting to produce an observation (Linacre, 1989, 1993). In the current context, there were two facets, examinee’s ability and rater’s severity interacting to produce the observed ratings. The goal was to estimate the examinee’s ability with the systematic rater effect removed. That is, the estimate of examinee abilities should be invariant across judges. The MFRM used in the scoring rubric analysis can be defined as,

$$\log\left(\frac{P_{irk}}{P_{irk-1}}\right) = \theta_i - R_r - F_k \qquad (5)$$

where P_{irk} denotes the probability of the examinee *i* judged by rater *r* to have a rating of k, the P_{irk-1} denotes the probability of the examinee *i* judged by rater *r* to have a rating of k-1, θ_i is the latent trait of the examinee i, the R_r is the severity of rater *r*, and finally F_k is the difficulty of rating step k relative to step k - 1.

In addition to fitting the MFRM model in equation 5, a model without estimation of the rater severity was also fitted (see equation 6) to examine the previous argument (e.g., Bramley, 2007; Linacre, 1989; Shah, Bradley, Parekh, Wainwright, & Ramchandran, 2013) that rater severity would be eliminated by design in the comparative judgment design. The MFRM used in the comparative judgment analysis was simplified to the rating scale Rasch model (Wright & Masters, 1982) as,

$$\log\left(\frac{P_{ik}}{P_{ik-1}}\right) = \theta_i - F_k \qquad (6)$$

where the parameters were defined in equation 5. The relative fit of the simplified model was examined to determine if ratings obtained via the comparative judgment method are less susceptible to systematic rater biases.

Model Estimation. The item response theory (IRT) models examined in this study were fitted using the Supplementary Item Response Theory Models (SIRT) package (Robitzsch, 2017) in R. The MFRM parameters can be estimated with maximum likelihood procedures. In the current study, the Marginal Maximum Likelihood Estimation (MMLE) was utilized since it permits the estimation of sample with extreme scores and short response strings (Linacre, 1999). The parameter estimates was obtained through a two-stage estimation approach, Expectation-Maximization (EM) algorithm (Bock & Aitken, 1981).

Latent Trait Estimation. The Expected A Posteriori (EAP) estimation was used in this study to obtain the latent trait estimates for each evaluated essay. The EAP estimation was derived from the Bayesian statistical principle and uses a prior probability distribution in conjunction of the response likelihood to estimate the posterior probability distribution of the latent trait scores. The prior probability was assumed to be normally distributed with a mean of 0 and standard deviation of 1.

Inter-rater Agreement. Intra-class-correlation (ICC) was calculated with the raw ratings provided by the raters. The ICC was considered to be one of the most commonly-used statistics for assessing inter-rater agreement (Hallgren, 2012). In contrast to Cohen's (1960) kappa coefficient that only accounts for perfect agreement, the magnitude of ICC incorporates the degree of disagreement among raters. In this study, ICC of a two-way mixed, absolute, average-measures model was estimated (McGraw & Wong, 1996) and defined as,

$$\frac{MS_{Essay} - MS_{E}}{MS_{Essay} + \frac{MS_{Rater} - MS_{E}}{n}} \tag{7}$$

where MS_{Essay} is defined as mean square for essay scores, MS_{Rater} is defined as mean square for rater scores, MS_{E} is defined as mean square error, and n is the number of essays. First, the two-way model was selected since each rating condition was fully crossed and accounted for rater variance. Second, absolute agreement is considered to be importance due to the claim that comparative judgment method can remove idiosyncratic interpretation of the rating scale (Böckenholt, 2004; Goffin & Olson, 2011). Third, average measures ICC was of interest since this study aimed to use the average ratings in each condition for hypothesis testing. Finally, the model was considered to be mixed since the raters of the study were considered as randomly selected from a population and the goal was to generalize the results to the population of possible raters.

In addition to point estimate, confidence intervals of the ICCs were also estimated and lower and upper limit are defined in equation 8 and 9 as,

$$\frac{n(MS_{Essay} - F^{*}MS_{E})}{F^{*}(MS_{Rater} - \ MS_{E}) + n(MS_{Essay})} \tag{8}$$

$$\frac{n(F^{*}MS_{Essay} - MS_{E})}{MS_{Rater} - \ MS_{E} + nF^{*}MS_{Essay}} \tag{9}$$

where F^{*} is defined as the F statistics associated with the available degrees of freedom and chosen alpha level. Also, n, MS_{Essay}, MS_{Rater}, and MS_{E} are defined in equation 7 above.

Estimated Trait Reliability. The reliability of the scores obtained from each evaluation condition could be calculated in a formula analogous to the classical test theory index of reliability. The formula involves calculating the ratio of true-score variance to observed score variance. In the current context, the reliability of the scores could be calculated by dividing the variance of the person parameter estimates by the variance of the latent trait (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Schroeders, Robitzsch, & Schipolowski, 2014) as follow,

$$\rho = \frac{Var(\theta)}{Var(\hat{\theta})} = \frac{\sigma^{2}(\theta_{EAP})}{\sigma^{2}} \tag{10}$$

where ρ denotes the reliability of the scores, $\sigma^{2}(\theta_{EAP})$ is the variance of the person parameter estimates, and σ^{2} is the variance of the latent trait.

3.1.6 Criterion-related Evidence of Validity

The criterion used for the criterion-related evidence of validity for this study was comprised of the latent trait estimates of the 100 essays used in the evaluation sessions. The latent trait estimates were obtained by fitting the MFRM model to the professional rater scores.

Additionally, ratings collected from each study condition were also fitted to the MFRM model to compute the latent trait estimates for each rating condition. Two metrics were used to evaluate the strength of the validity evidence of each scoring method. First, Pearson correlation was used to examine the relation of the latent trait estimates obtained from each study condition and the gold-standard. Second, root-mean-squared-error (RMSE) was computed for each rating condition to evaluate the degree of deviation from the gold-standard.

3.1.7 Mixed-Design ANOVA

The self-reported ease of use, enjoyment, and time taken to evaluate each essay were analyzed using the mixed-design ANOVA. Since self-reported data was only collected at the end of each evaluation session, the model treated the session number as the within-subject variable and rating condition as the between-subject variable. For the time taken dependent variable, essay number was treated as the within-subject variable with rating condition remained to be between-subject variable.

Chapter 4

4.1 RESULTS

Purpose of this chapter is to provide the analysis results of the effect of different rating conditions. This chapter starts with descriptive statistics of the participants who partook in the training sessions, and essays that were used in the evaluation sessions. Second section of this chapter describes the results of the training session and provides the demographic information of the participants who passed the training session. Third, the five proposed hypotheses were tested. Finally, the chapter ends with a summary of the hypothesis testing results.

4.1.1 Descriptive Statistics

Participants Characteristics. All participants recruited in this study were U.S. residents. The participants age varied from 22 to 63 years (*M* = 34.5, SD = 11.4), 38% were women, most had at least some postsecondary education (7% were high school graduates, 18% had associate degree, 25% some college, 45% bachelor degree, 4% graduate degree), teaching experience varied from 0 to 9 years (*M* = 0.55, SD = 1.55), and had some grading experience (*M* = 1.87, SD = 0.84).

Essay Characteristics. The 1785 essays, on average, had 350 words and received a score of 4.26 (SD = 0.77). The average score was calculated as the average of the two professional rater scores and no third rater score was needed to resolve any agreement. The essay scores were positively skewed with relatively little essays received scores lower than 2.5 (see figure 4.1). In comparison, the 100 essays used in the evaluation sessions were more normally distributed (*M* = 3.85, SD = 1.1; also see figure 4.2). The descriptive statistics of the rating distributions can be found in table 4.1. Also, a subset of the data used in this study can be found in Appendix H.

Table 4.1. Descriptive Statistics of Expert Ratings

	N	*M*	S.D.	Skewness	Kurtosis
Gold-Standard (1785)	1783	4.26	0.77	-0.46	1.66
Gold-Standard (100)	100	3.85	1.1	-0.35	-0.09

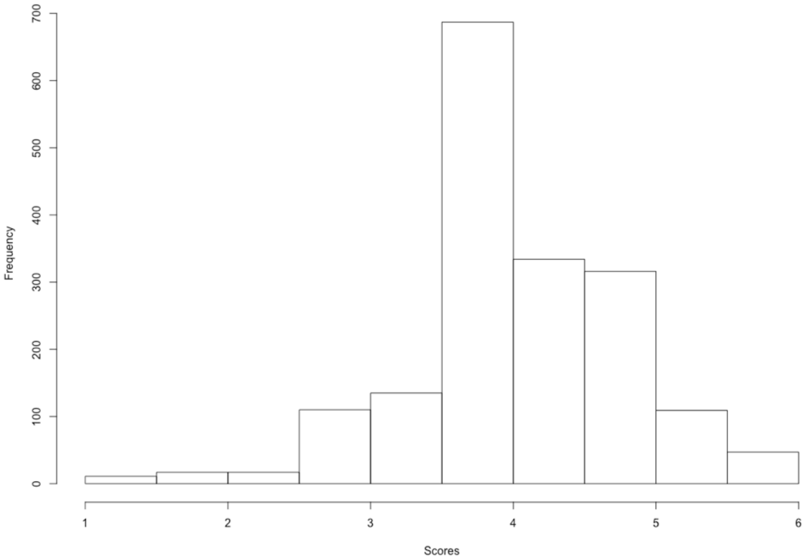

Figure 4.1. Distribution of All Essay Scores.

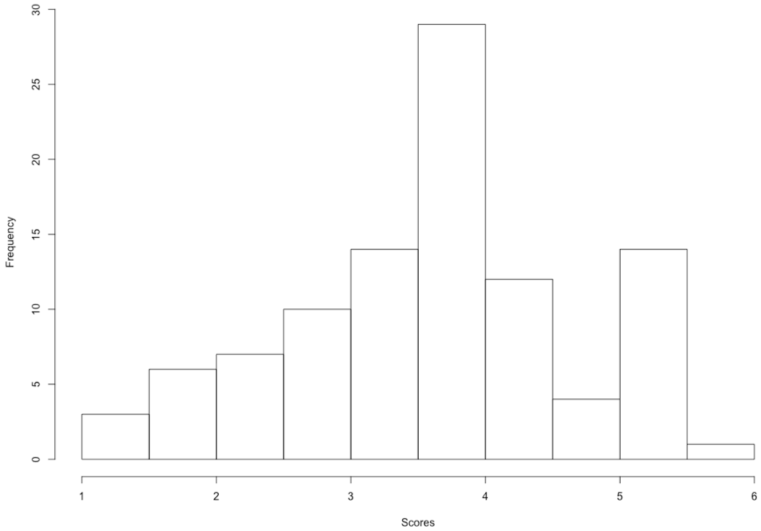

Figure 4.2. Distribution of Evaluation Essay Scores.

4.1.2 Training Results

At the beginning of the study, recruited participants were randomly assigned to one of the training conditions in the beginning of the study. However, it was found that the participants in the two authentic rubric conditions were able to pass the training sessions at a higher rate. As a result, more raters in the authentic rubric conditions were completing the evaluation sessions. As an effort to maintain a balanced study design, all participants were assigned to the rating condition after a sufficient number of participants were recruited for the two authentic rubric conditions. Therefore, the total number of participants assigned to the rating condition was greater than either one of the authentic rubric condition.

The training session success rate can be found in table 4.2. Among the 22 participants assigned to the rating condition, 12 of them could achieve the minimum performance standard, resulting in a passing rate of 54.5%. In contrast, the two authentic rubric conditions were able to achieve higher passing rate (73.3% and 61.1%, respectively). A chi-square test of independence, however, revealed no significant relation between evaluation condition and number of participants completing the training session ($\chi^2{}_{df=2}$ = 1.340, p-value = 0.512).

Table 4.2. Training Results Across Conditions

	Rating	AR_{Odd}	AR_{Even}
Passed	12	11	11
Failed	10	4	7
Total (%)	22 (54.5%)	15 (73.3%)	18 (61.1%)

Note. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

Descriptive Statistics. Among the 34 participants who passed the training session, 22 completed all five evaluation sessions. All the following analyses only utilize the data from these 22 participants. The demographic information of the raters who completed all five evaluation sessions are presented in table 4.3. The demographic makeup of the raters in all three conditions were fairly similar to all the recruited participants. Comparing among the evaluation conditions, raters in the rating condition were younger and reached lower education level, but had more teaching experience. Raters in the odd scores authentic rubric condition were the oldest, predominantly male, and had no teaching experience.

Table 4.3. Demographic and Self-Report Data of the Three Study Conditions

	Rating (n = 6)	AR_{Odd} (n = 7)	AR_{Even} (n = 9)
Age	28.17 (7.33)	38.29(14.72)	35.44(10.24)
Gender (Female)	50%	14.3%	44.4%
Education Level			
High School Graduate	16.7%	0%	0%
Associate Degree	50%	14.3%	11.1%
Some College	16.7%	28.6%	33.3%
Bachelor Degree	16.7%	57.1%	55.6%
Graduate Degree	0%	0%	0%
Teaching Experience	1.17(2.04)	0.00(0.00)	0.22(0.44)
Grading Experience	2.33(1.21)	1.43(0.53)	2.33(0.87)
Evaluation Duration(s)	15.987(15.197)	17.638(17.000)	16.010(15.145)
Ease of Use	4.033(1.520)	4.114(1.367)	3.556(1.120)
Enjoyment	1.667(0.844)	2.086(1.173)	2.933(1.176)

Note. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

4.1.4 Model Fit

To examine if a model without systematic rater effects would fit the authentic rubric rating, the two models described in equation 5 and 6 were fitted to the evaluation data collected from the two authentic rubric conditions. For odd scores rubric condition, the model with rater parameters estimated provided statistically significant better fit with the evaluation data (${\chi^2}_{df=6}$ = 189.414, p-value < 0.001). Similarly, the more complex model was found to provide a better fit with the even scores authentic rubric condition evaluation data. (${\chi^2}_{df=8}$ = 153.106, p-value < 0.001). In summary, MFRM with rater effects estimated was found to provide significantly better fit with the rating data collected via the use of authentic rubric (see table 4.4). Therefore, the rater effect model would be used to fit the model and estimate the latent trait in the following analysis and rater parameter estimates are presented in table 4.5. Examining the rater severity estimates reveals that there were substantial variations in systematic rater effects in each study condition. The difference between the most stringent and most lenient raters were estimated to be 1.687, 3.076, and 1.677 logits for holistic rubric, odd scores authentic rubric, and even scores authentic rubric, respectively. The substantial systematic rater differences also cast doubt onto the argument that comparative judgment method can eliminate systematic rater biases by design.

Table 4.4. Model Fit Information

Model	AIC	BIC	Deviance	No. of Parameters	χ^2	df	p-value
Rating							
No rater effect	1455.420	1473.656	1441.420	7	78.652	5	<.001*
Rater effect	1386.768	1418.030	1362.768	12			
AR_{Odd}							
No rater effect	1643.598	1661.835	1629.598	7	189.414	6	<0.001*
Rater effect	1466.187	1500.051	1440.184	13			
AR_{Even}							
No rater effect	2375.883	2394.119	2361.883	7	153.106	8	<0.001*
Rater effect	2238.776	2277.854	2208.776	15			

Note. * $p < 0.05$; AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

Table 4.5. Rater Parameter Estimates

Condition	Rater	Severity Measure	SE
Rating	1	0.364	0.159
	2	0.181	0.159
	3	-0.012	0.160
	4	-1.029	0.164
	5	0.658	0.158
	6	-0.161	0.161
AJ_{Odd}	1	-0.684	0.177
	2	-1.295	0.179
	3	-0.020	0.177
	4	0.340	0.177
	5	0.311	0.177
	6	1.781	0.179
	7	-0.433	0.177
AJ_{Even}	1	0.643	0.145
	2	-1.034	0.140
	3	-0.956	0.140
	4	0.622	0.145
	5	-0.360	0.142

6	-0.098	0.142
7	0.166	0.143
8	0.393	0.144
9	0.626	0.145

Note. * $p < 0.05$. Rating: Holistic rubric; AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

4.1.5 Hypothesis Testing

The data collected from the evaluation sessions are used to examine the five proposed hypotheses.

Hypothesis 1: It was proposed that the rater evaluations collected via the authentic rubric procedure will have higher reliability than those collected using traditional rubrics. The reliability of the rater evaluations was analyzed through two metrics: score reliability and intra-class correlation (ICC). In the case of score reliability, ratings collected from all three conditions were found to have high reliability (see table 4.6) indicating that the EAP estimates was a consistent estimate of the latent trait.

In addition to the score reliability, the ICC provided a means to assess the degree that raters provided consistency in their ratings of essays. For each study condition, a two-way mixed, absolute, average-measures of ICC (McGraw & Wong, 1996) was estimated. The estimated ICCs for the rating, CJ1, and CJ2 conditions were 0.698, 0.737, and 0.716, respectively (see Table 4.6). All three estimates fell in the range that would be considered as indicating good degree of agreement (Cicchetti, 1994). Although the two authentic rubric conditions yielded marginally higher ICC estimates, closer examination of the confidence intervals reveals that there was substantial overlapping in the range of estimates, indicating that the ICC estimates were not significantly different from each other. Finally, it is worthwhile to note that the estimated inter-rater agreements, while could be considered to be in the good range, were significantly lower than the two professional raters (ICC = 0.879).

Table 4.6. Intraclass Correlation Coefficients and Score Reliability

Study Condition	ICC (95%CI)	EAP Reliability
Gold-Standard	0.879 (0.819, 0.918)	0.797
Rating	0.698 (0.613, 0.773)	0.945
AR_{Odd}	0.737 (0.642, 0.812)	0.962
AR_{Even}	0.716 (0.642, 0.783)	0.964

Note. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

In addition to examining score reliability and ICCs, the sizes of the standard errors of the estimates were also evaluated to determine if the authentic rubric condition provided more precise measurement. In this case, we found that the rating condition and the odd scores condition have similar standard error magnitude, whereas the even scores authentic rubric condition has the lowest standard error size (see table 4.7).

Table 4.7. Standard Error Sizes

	Mean(SD)
Rating	0.68(0.07)
AR_{Odd}	0.68(0.06)
AR_{Even}	0.48(0.09)

Note. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

Hypothesis 2: It was hypothesized that the authentic rubric procedure will result in scores that have stronger evidence of validity than the traditional rubric. Specifically, it was expected that authentic rubric scores will have higher degrees of agreement with professional rater scores. In this study, two estimates of professional rater scores were used. The two estimates include 1) an MFRM latent trait estimate based on all 1785 essays in the original dataset (Gold-Standard (1785)), and 3) an MFRM latent trait estimate based on the 100 essays (Gold-Standard (100)) used in the evaluation sessions. For each study condition, the MFRM estimates were reported.

Although it was originally expected that fitting the MFRM with the full essay sets would yield more accurate rater effect estimates, examining the correlation matrix (see table 4.8) revealed that the MFRM estimates obtained using full essay sets (1785) and essays used in the evaluation sessions (100) were highly correlated (r = .991). Therefore, it was not surprising to find that, regardless of whether full or subset of the gold-standard essay scores were used in fitting the MFRM, the correlation pattern between MFRM estimates obtained from each study condition and gold-standard estimates was consistent. Specifically, it was found that the even scores authentic rubric condition had the highest correlation with the gold-standard estimates (r = .880 & .900). Odd scores authentic rubric condition had the second highest correlations with the gold-standards (r = .873 & .888), while rating condition produced the lowest correlation (r = .840 & .857). The difference in correlation magnitudes between even scores authentic rubric and rating conditions would be considered small to medium (Cohen's q = .155 & .190; Cohen, 1992). These results suggest that the authentic rubric conditions, in comparison to the use of holistic rubric, yielded evaluations that have the highest accuracy. However, significance tests using fisher transformation revealed no significant difference in the correlation magnitude among the three conditions. For instance, comparing the even scores authentic rubric condition and rating condition correlations found that the difference was not statistically significant (z = 1.33, p = 0.09).

Table 4.8. Correlations Among MFRM Estimates

	1	2	3	4
1. Gold-Standard (1785)				
2. Gold-Standard (100)	0.991			
3. Rating estimate	0.840	0.857		
4. AR_{Odd} estimate	0.873	0.888	0.957	
5. AR_{Even} estimate	0.880	0.900	0.955	0.962

Note. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

In order to quantify the error magnitude of rater scores from each study condition, RMSEs between MFRM estimates of each study condition and professional rater scores were computed (see Table 4.9). The results had two characteristics that are worthwhile to note. First, using all the essays in the dataset to fit the MFRM model resulted in lower RMSE. The difference in RMSE was as large as 1.0, where range of latent trait estimates is approximately 14 logits. Second, the degree of correspondence between rater estimates and the gold-standard is consistent regardless of the data source used. Specifically, the even scores authentic rubric condition yielded lowest RMSEs, the rating condition had the second lowest RMSEs, and the odd scores authentic rubric condition has the highest RMSEs. Compared to the rating condition, the even scores authentic rubric condition reduces RMSEs by 7.2% to 13.5%. The rating condition, in turn, has 1.3% to 7.5% smaller RMSEs than the odd scores authentic rubric condition.

Table 4.9. RMSE of Each Study Condition

	Gold-Standard(100)	Gold-Standard (1785)
Rating	1.520	2.413
AR_{Odd}	1.568	2.361
AR_{Even}	1.278	2.292

Note. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

Hypothesis 3 and 4: It was hypothesized that the raters will find the authentic rubric to be both easier and more enjoyable to use. Examining the self-reported ease of use descriptive statistic in table 4.3 showed that the data was trending towards the opposite direction. Specifically, raters found holistic rubrics to be the easiest to use, whereas even scores authentic rubric was reported to be the most difficult one to use. Nonetheless, analyzing the self-reported ease of use data with mixed-effect ANOVA model revealed that there was no statistically significant difference in self-reported ease of use between the study conditions ($F_{2,103}$ = 2.114, p

= 0.126). Since the omnibus test failed to reach statistical significance, no further follow-up test was conducted.

Although there was insufficient evidence of difference in ease of use between study conditions, an examination of self-reported enjoyment data revealed statistically significantly differences between the study conditions ($F_{2,103}$ = 13.209, p < 0.001). Follow-up test (see table 4.10) showed that raters in the even scores authentic rubric condition reported significantly higher enjoyment than both the rating and the odd scores authentic rubric condition. In particular, the difference between the even scores authentic rubric condition and the rating condition (0.952) was greater than the difference between the even scores authentic rubric condition and the odd scores authentic rubric condition (0.579).

Table 4.10. Follow Up Test of Self-Reported Enjoyment

	Estimate	SE	Z	p-value
AR_{Odd} vs. Rating	0.373	0.238	1.572	0.257
AR_{Even} vs. Rating	0.952	0.220	4.327	< 0.001*
AR_{Even} vs. AR_{Odd}	0.579	0.223	2.595	0.026

Note. * p < 0.05. AR_{Odd}: Odd scores authentic rubric; AR_{Even}: Even scores authentic rubric

Hypothesis 5: The last hypothesis proposed that the use of authentic rubric can be more efficient than using traditional holistic rubric. In this case, the efficiency of the evaluation procedure was assessed by the amount of time it took to evaluate each essay. Using the one-way mixed effect linear model, it was found that the difference in evaluation time among study conditions did not reach statistically significance ($F_{2,2098}$ = 2.576, p = 0.076). Therefore, no follow-up test was conducted to examine the differences in efficiency between study conditions.

4.1.6 Conclusion

Five hypotheses were tested in this study. Based on the above results, one of the five hypotheses was supported. Specifically, using mixed-effect ANOVA model, it was found that the raters in the even scores authentic rubric condition reported statistically significantly higher enjoyment than both the rating condition and odd scores authentic rubric condition. In other words, raters who used essays that received scores of 2, 4, and 6 as anchors found the task to be more enjoyable than the raters who used holistic rubric or essays that received scores of 1, 3, and 5.

Although authentic rubric condition was found to be more enjoyable to use, no evidence was found to support authentic rubric producing more reliable, valid scores or being easier to use or more efficient. Examination of ICCs found that both authentic rubric conditions had marginally higher inter-rater consistency. The difference in ICCs between the authentic rubric conditions and rating condition, however, failed to reach statistical significance. Examination of score reliability and standard error of estimates portrayed a similar story where odd scores authentic rubric condition did not differ substantially from the rating condition while even scores authentic rubric condition had the highest measurement precision. In terms of validity of the scores, it was found that scores from both authentic rubric conditions had marginally higher correlations with the professional rater scores, with the even scores authentic rubric condition had the highest correlation. Nonetheless, the difference in correlation magnitude did not reach statistical significance. In addition to examining correlations, the RMSE of the latent trait estimates for each study condition was also computed. Similar to the correlation result, even scores authentic rubric condition was found to have highest degree of agreement with the gold-standard estimates, which translated to lowest RMSE. Rating condition, however, had the second lowest RMSE. Finally, using one-way mixed effect linear model found no evidence that the study conditions statistically differed on either self-reported ease of use and amount of time it took to evaluate each essay.

Chapter 5

5.1 PURPOSE OF THIS STUDY

This dissertation aims to study the feasibility of combining holistic rubric with comparative judgment procedure to retain the theoretical advantages of comparative judgment methods while overcoming its weaknesses. The proposed authentic rubric conceptualized holistic rubric as a series of comparative judgments against pre-selected anchors. Five hypotheses were tested in this dissertation to examine if the proposed authentic rubric would have stronger psychometric properties and provide greater rater experience. Particularly, it was examined if authentic rubric will elicit more reliable and valid rater judgments, while being easier, more enjoyable, and more efficient to use. Participants were recruited from the Mturk platform and assigned to one of the three study conditions to use either holistic rubric or one of the two authentic rubrics to evaluate 100 essays. The psychometric properties of the rater judgments and self-report data collected from the participants were used to test the five hypotheses. This chapter discusses the findings of this study and its implications for future performance assessment evaluation procedure. Also, limitations of the study and important areas for future research are also highlighted.

5.2 DISCUSSION OF THE OVERALL FINDINGS

The first hypothesis, which proposed that rater judgments elicited via authentic rubric method would have higher reliability and inter-rater agreement than holistic rubric, was not supported. Using ICC, the two authentic rubric conditions were found to have marginally higher inter-rater agreement than the holistic rubric condition. Nonetheless, a closer examination revealed that there was substantial overlap in the confidence intervals indicating that the magnitude differences were not statistically significant. Examination of the ICCs and standard error magnitude revealed a similar trend where the even scores authentic rubric condition had marginally higher reliability and precision of measurement. This finding was similar to Attali's (2014) report that comparative judgment method, ranking in this case, elicited rater judgments that had higher consistency than holistic rubric judgments. However, Attali (2014) did not examine if the differences were statistically significant. This study showed that the interrater agreement of comparative judgment methods may not be statistically significantly higher than judgments obtained via holistic rubrics. In addition to being statistically nonsignificant, the differences in ICCs magnitude were also fairly small with authentic rubrics having .018 to .039 higher ICCs. Both the significance test and practical significance would suggest that comparative judgment method failed to promote higher inter-rater agreement. The findings raised question regarding one of the theoretical advantages of comparative judgment methods. Advocates argued the use of comparative judgment methods would remove the arbitrary nature and idiosyncratic implementation of rating scales (Böckenholt, 2004; Goffin & Olson, 2011), this study, however, did not find sufficient empirical evidence to support such claim. The introduction of common anchors did not improve consensus on evaluation criteria substantially.

The second hypothesis, which suggested that rater judgments elicited via the authentic rubric method would have stronger evidence of validity than holistic rubric, was not supported. To examine the validity of the scores, rater judgments were compared to professional rater scores through correlations and comparisons of magnitudes of RMSEs. Although the correlation pattern was found to favor authentic rubric methods, significance tests of correlation magnitudes did not reveal any statistically significant difference among the study conditions. Finally, examining the RMSEs found that the even scores authentic rubric condition had the lowest RMSE value and a modest reduction in RMSE in comparison to the rating condition.

It is worthwhile to note that the authentic rubrics were comprised of 7 score points whereas the holistic rubric had 6 score points. The disparities in scale length could have accounted for the correlation magnitude differences. In order to address this possible confound, all scores collected via authentic rubrics were transformed to a 6-point scale. For the odd scores authentic rubric, essays that were judged to be worse than anchor 1, which corresponds to a score of 0, were rescored to a score of 1. At the same time, essays that were judged to be superior than anchor 3 in the even scores rubric condition were rescored to receive a score of 6. In other words, both authentic rubric conditions had a minimum score of 1 and maximum score of 6, which is equivalent to the holistic rubric. Correlations between the gold-standards and the two authentic rubric condition rescored ratings were then estimated. In both cases, the rescoring did not result in substantial changes in correlation magnitudes. The order of the correlation magnitudes remain to be the same among the three study conditions.

Although the current study did not find sufficient statistical evidence to support the validity of comparative judgment method, it was largely consistent with the existing comparative judgment literature that found comparative judgment method often yielded the same, if not higher, correlation with the gold-standard. In this case, the difference in correlations magnitudes would be considered as small to medium (Cohen, 1992). One plausible reason for the statistically nonsignificant finding is that the current study had relatively small sample size. Post-hoc analysis revealed that the current study had limited statistical power ($1 - \beta = .375$). Future study that aims to replicate the result with sufficient statistical power ($1 - \beta = .95$) should strive for having raters in each condition rating at least 600 essays.

Beside the limited statistical power, another possible explanation for the statistically nonsignificant results found in this study may stem from the limitation of the materials.

Particularly, previous studies that reported comparative judgment methods yielded higher criterion evidence of validity either used some form of external criteria such as prior achievement (McMahon & Jones, 2014), physical measurements (Shah et al., 2014), and cognitive test performances (Goffin et al., 1996); or expert evaluations that involved multiple experts (Sheppard et al., 2011; Wagner & Goffin, 1997). In this study, the validity evidence available was limited to scores on a 6-point holistic rubric provided by two professional raters. In other words, the range of possible scores was limited to 12 points which could only approximate the latent distribution of essay qualities. Thus, it was possible that the authentic rubric was more capable than the holistic rubric at recovering the latent distribution; however, this capability could not be demonstrated with the limited range and variance of the gold-standard scores.

Finally, it is worthwhile to point out that the current study implemented a training and selection procedure that was shown to be quite effective in coaching relatively inexperienced raters (Attali, 2015). After going through the training session, Attali found that inexperience raters were able to evaluate essays as reliably and validly as expert raters. In this case, it was not surprising to find that raters in all three conditions provided evaluations that had good psychometric properties. Implementation of such training procedure, however, may not be feasible in all performance assessment context. For instance, peer assessment in classrooms and MOOCs often do not involve comprehensive training procedure due to concern of training workload imposed upon the peer raters. As a result, peer raters often received little to no training on the evaluation task. The lack of training would, as expected, result in degrading of the psychometric properties of the evaluations. Although this study did not directly examine the psychometric properties of authentic rubric methods with novice raters who received minimal training, the training session data collected in this study shed some light on the potential advantageous of the comparative judgment methods. Specifically, it was reported that the raters enrolled in the two authentic rubric conditions were able to pass the training requirement at a higher rate than the participants in the holistic rubric condition. This result suggests that, using authentic rubrics, novice raters who received minimal training produced evaluations that had high agreement with the expert raters. In comparison, participants who used holistic rubric struggled to adopt the expert evaluation standards prior to the completion of the training procedure. These results indicate that when comprehensive training is not feasible, utilizing authentic rubric may promote higher quality rater judgments.

Third and fourth hypotheses, which stated that authentic rubric would be easier and more efficient to use than holistic rubric, was not supported. Examining the descriptive statistics revealed that the self-reported ease of use data actually went against the hypothesis that the even scores authentic rubrics were found to be the most difficult to use. However, mixed-design ANOVA analysis of the data found no statistically significant difference among the study conditions in the perceived ease of use. Furthermore, mixed-design ANOVA analysis of the time used to evaluate each essay revealed no statistically significant difference among each study conditions. The findings were surprising in two aspects. First, proponents of comparative judgment methods argued that comparative process is inherent in all evaluations (Goffin & Olson, 2011). Past work on studying rater cognitive processes also found that comparison between pieces of essays were common in writing assessments (Crisp, 2012). Additionally, it was argued that raters encode training materials into a mental scoring rubric (Bejar, 2012) which represents their internalized understanding of the scoring criteria and their mental representations of examples of work they have seen (Crisp, 2012). Thus, it was expected that the mental representations of the holistic rubric, which consists of six explicit score points, would find to be more mentally complicated than the authentic rubrics, which was composed of three anchors. The combination of a scoring procedure that aligns with the underlying cognitive processes and lessens the mental burden imposed upon raters’ mental resources was expected to reduce the mental workload involved in the evaluation process. Although the raters who used the odd scores authentic rubric reported marginally higher ease of use than raters who used the holistic rubric, the difference was too small to reach either statistical and practical significance. In other words, empirical evidence did not provide support to the claim that comparative judgment and smaller sets of anchors would reduce the mental workload of the raters.

The second surprising aspect of this finding is that the differences in reported ease of use, albeit not reaching statistical significance, did not translate to differences in time taken to evaluate essays. The average time taken to evaluate essay differed by less than a second between each condition. Since the anchors used in the even scores authentic rubric were of higher quality and considerably longer (see Appendix D) than the ones used in the odd scores authentic rubric, it was therefore expected that raters who used the even scores authentic rubric would had to memorize more information which would cause the internalization process of the anchors to be more difficult. As a result, when and if the raters needed to reference the anchors in the rubric, raters in the even scores authentic rubric condition were expected to take more time and effort to read the anchors. The fact that the high authentic rubric condition raters did not spend substantially more time on evaluating essays suggest that the raters in the authentic rubric conditions spent minimal time on reviewing the anchors during evaluations.

Plausible explanation for these findings was that anchors in the authentic rubric were not perceived and internalized as individual essays. Rather, raters internalized the patterns and features of the anchors and used the same criteria for evaluation (Crisp, 2012). In this case, the raters in the even scores authentic rubric condition may initially require additional effort in studying the anchors. However, the mental effort required during evaluation would be substantially lower since only partial features of these anchors would be internalized by the raters. In other words, the mental effort required for evaluation would not be substantially different across study conditions once raters were able to complete the training session. Whether raters did engage in the proposed cognitive processes would require future study to collect additional empirical evidence.

The last hypothesis, which proposed that raters would find the authentic rubric to be more enjoyable to use than holistic rubric, was supported. Analysis of the self-reported enjoyment at the end of each evaluation session using mixed-design ANOVA model found that the raters in the even scores authentic rubric condition reported statistically significantly higher enjoyment. Follow-up post hoc analysis revealed that the raters in the even scores authentic rubric condition reported higher enjoyment than the raters in the holistic rubric condition, whereas raters in the odd scores authentic rubric and holistic rubric conditions reported similar level of enjoyment. While it was possible that the raters in the even scores authentic rubric condition reported higher enjoyment due to the novelty effect of using an evaluation procedure differ from conventional practice, raters in the odd scores authentic rubric condition used a same novel procedure and did not report statistically significant higher enjoyment than raters in the rating condition. The results indicated that the even scores authentic rubric condition, albeit required reading and comparing against anchors of longer length, presented a more enjoyable task for the raters. One possible explanation for this finding was that tasks with optimal challenge could have a positive effect on the quality of the experience (Csikszentmihalyi & LeFevre, 1989; Moneta & Csikszentmihalyi, 1996). In this study, reading and comparing against the longer and higher quality anchors could present a more challenging task (Wolfe, Song, & Jiao, 2016), which, in turn promoted the evaluation experience. In contrast, raters in the odd scores authentic rubric condition may find the tasks to be too easy to complete. In particularly, the anchor that received a score of 1 is incomplete and contains very little information (see Appendix C), which would require considerably less effort to compare against. The overall lower degree of challenge presented by odd scores authentic rubric would therefore degrade the raters' experience.

5.3 IMPLICATION FOR PERFORMANCE ASSESSMENT

Developing a holistic rubric requires effort in identifying and operationalizing the constructs intended to assess, deciding the length of the scale, and developing descriptors for each scale point. Each of these components requires extensive attention from the assessment developer. In a review of 21 articles published in guiding rubric design, Tierney and Simon (2004) found that most published guidelines focused on clarity of the descriptors and its impact on reliability of the interpretations made by the raters. Others also stressed the importance of descriptions in differentiating scale points (Moskal, 2003; Wiggins, 1998), and improving usability and reliability (Popham, 1997; Wiggins, 1998). The findings of this study suggest that effort expended upon developing detailed and accurate descriptors of score levels may not be necessary. Substituting the score descriptors with a small set of anchors resulted in little loss in validity and reliability of the assessment while improving raters' experience under certain condition. As a result, the process of developing a psychometrically sound rubric can be greatly simplified. In addition to streamlining the rubric development process, the findings in this study can also inform the rangefinding practice in the construction of holistic rubrics. Specifically, the current study demonstrated that a smaller number of benchmark performances than recommended (e.g., Johnson, Penny, & Gordon, 2009; Livingston, 2009; Wolcott & Legg, 1998) were needed to operationalize the evaluation criteria. The authentic rubrics in this study used only half of the traditional number of benchmark performances while suffering no loss in interrater agreement and validity.

5.4 FUTURE RESEARCH

The findings of this study demonstrated that it is possible to simplify the existing holistic rubric to a series of comparative judgments. In this case, an authentic rubric with three anchors was enough to match and, in some cases, exceed the performance of a holistic rubric. The possibility of conceptualizing a holistic rubric as a series of comparative judgments against anchors would introduce a new area of research. First, future research can investigate what are the best guidelines in constructing the authentic rubrics. For instance, authentic rubric used in this study utilized three anchors to mimic a 7-point rating scale. If the authentic rubric used one anchor to illustrate each score point of the holistic rubric, the authentic rubric in this study would approximate a 13-point rating scale. Previous research found that increasing score points in the rubric have a diminishing return with reliability estimate (Shumate, Surles, Johnson, & Penny, 2007) and raters may underutilize certain scoring categories (Johnson et al., 2009). At the same time, past research (Attali, 2014) found that raters were capable of reliably ranking 5 performances at once. Therefore, it is unclear if the use of additional anchors in the authentic rubric would also yield diminishing return on measurement precision, or, if raters are capable of comparing more anchors at the same time.

In addition to the number of anchors used in the rubric, future research can also investigate the characteristics of anchors that may result in best discrimination in scoring. One surprising finding from this study is that the use of longer and higher quality essays as the anchors resulted in higher inter-rater agreement, higher correspondence with the gold-standard scores, and higher enjoyment with no loss on efficiency nor perceived ease of use. Although all of the aforementioned findings did not reach statistical significance, it may indicate a trend that comparison against more sophisticated essays can yield higher quality rater judgments. Future research can therefore investigate what characteristics of the anchors may impact the quality of the rater judgments.

Second, the creation and selection of anchors in this study is based on professional rater scores collected via holistic rubric. Therefore, the distinction between each essays are confined to the 6-point scale used by the raters. It is of interest to investigate if professional raters used a comparative judgment procedure, such as ranking, to evaluate a subset of essays could provide higher discrimination among the essays. The increased discrimination among essays can inform the selection of anchors and improve the performance of the comparative judgment procedures. The alignment of the anchors selection process and the evaluation procedure may also result in an improvement of the evaluation results.

Third, the current study implements the comparative judgment methods in a fashion that is similar to the holistic rubric format. However, the educational assessment literature on comparative judgment methods often implement the forced choice format where tie is not allowed. In the context of this study, raters will only be offered the choices of deciding whether the essays to be evaluated are better or worse than the anchors without the option of deciding that they are as good as the anchors. The forced choice format can theoretically eliminate the systematic rater biases through the rating design. However, whether such claim is warranted requires empirical evidence to support. Furthermore, it will be important to examine if raters would be tolerant to such forced choice format.

Fourth, although the current study failed to find substantial evidence that the authentic rubric condition had superior psychometric properties than the holistic rubric, the current study did find that authentic rubric is at least as capable as holistic rubric at collecting reliable and valid evaluations from raters. Instead of focusing on demonstrating the psychometric properties of the authentic rubric, future studies should explore the theoretical flexibility of the authentic rubric format. For instance, removing the reliance on numerical scores would allow each rater to use different anchors to evaluate the responses. Each rater in this case would provide different amount of information at various points of the measurement scale. This suggested design would allow researchers to develop an adaptive evaluation system that is similar to the design created by Pollitt (2012).

Finally, as previously discussed, the training procedure implemented in this study eliminated a substantial amount of rater variations and authentic rubric may be proven to have most utility value in situation where comprehensive training of raters may not be feasible. This suggestion was based primarily on the small number of ratings collected during the training session. In order to verify the utility value of authentic rubric for peer assessments in classrooms and MOOCs, it is necessary for future research to examine if the authentic rubric method is more robust against rater variations that are more commonly observed when training is not provided.

5.5 LIMITATIONS

There are several limitations of this study. First, as previously mentioned, the validity evidence available was limited to the two professional rater scores. In this dataset, the expert raters are permitted to differ by 1-point on a 6-point scale. In other words, there was substantial variance among the expert ratings. With the professional rater scores also subjected to random errors and rater biases, the reliance on one source of validity evidence both limited the ability to thoroughly investigate the accuracy of the evaluation procedure and the generalizability of the results. In order to examine if the results of this study can be generalized into other context, it is necessary for future studies to collect additional form of validity evidences on multiple forms of performance assessment.

Second, this study examined the psychometric properties of the use of authentic rubric with one form of essay structure using one dataset. The essays used in this study were relatively short and easy to read. It requires additional studies to examine if the authentic rubric format can be used with essays of written to different formats and different length. For instance, for essays that are written at a higher grade levels, the length of the essays was usually longer, which in turn, requires the raters to invest more time into reading and internalize the anchors. Whether current study findings will be generalizable to such scenario still requires further investigation.

Third, the current study employed the evaluation procedure via the internet with minimum interaction among raters and between raters and the researcher. Large-scale performance assessments often implement procedure to monitor rater accuracy and recalibrate the results throughout the evaluation sessions as a safeguard against raters drifting away from the consensus standard. Theoretically, these procedures would ensure greater agreement among raters and higher quality rater judgments. However, due to the limitation of the design of this study, these safeguard procedures could not be implemented. Future studies will need to examine how these traditional procedures interact with the use of authentic rubric and what modifications may be needed to suit the new evaluation format.

5.6 CONCLUSION

Although this study has several limitations, the findings of this study showed that it is feasible to combine the features of the comparative judgment with holistic rubric scoring by replacing score points with anchor performances. The new evaluation format was proved to have comparable psychometric properties to the holistic rubric evaluation format while improving the rater experience during evaluations. Empirical evidence also suggested that proposed evaluation format could promote slightly higher interrater agreement and criterion-related correlations. Within the Kane's (1992, 2013) argument-based validation framework, the current study showed that there is sufficient empirical evidence to support the use of authentic rubric as a scoring procedure in performance assessment. Nonetheless, some of the current study findings cast doubts on the claims of comparative judgment theoretical advantages. In particular, the claim that comparative judgment method can eliminate systematic rater biases was rejected by model fit testing. With relatively little empirical evidence available in comparative judgment educational assessment, there is a need for future research effort to continue investigate which, if any, theoretical advantages of comparative judgment would be supported by empirical data. Finally, this study introduced the opportunity to streamline the procedure of developing holistic rubric and save costs in the process. The change in the holistic rubric format can, as argued previously, potentially enable new form of evaluation procedure that may not be feasible with holistic rubric. This study also highlighted the importance of further our understanding of the cognitive processes engaged by the raters during the evaluation sessions. The understanding is quintessential in the future effort to improve the quality of rater evaluations in the future.

References

Alharby, E. R. (2006, May). *A comparison between two scoring methods, holistic vs analytic, using two measurement models, the generalizability theory and the many-facet rasch measurement, within the context of performance assessment*. Pennsylvania State University, University Park, PA.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, *74*(5), 795–808. https://doi.org/10.1177/0013164414527450

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, 0265532215582283.

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141.

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*(12), 1210–1218. https://doi.org/10.1037/0003-066X.48.12.1210

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, *27*(4), 515–535. https://doi.org/10.1177/0265532210368717

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bissell, A. N., & Lemons, P. P. (2006). A new method for assessing critical thinking in the classroom. *BioScience*, *56*(1), 66–72. https://doi.org/10.1641/0006-3568(2006)056[0066:ANMFAC]2.0.CO;2

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, *9*(4), 453–465. https://doi.org/10.1037/1082-989X.9.4.453

Böckenholt, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika*, *71*(4), 615–629. https://doi.org/10.1007/s11336-006-1598-5

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item paramters: An application of the EM algorithm. *Psychometrika*, *46*, 443–459.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, *14*(4), 21–24. https://doi.org/10.1111/j.1745-3992.1995.tb00885.x

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 264–294.). London: QCA: Eds.

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, *25*(2), 1–35.

Caragiannis, I., Krimpas, G. A., & Voudouris, A. A. (2014). Aggregating partial rankings with applications to peer grading in massive online open courses. *arXiv Preprint arXiv:1411.4619*. Retrieved from http://arxiv.org/abs/1411.4619

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24*(4), 310–324. https://doi.org/10.1177/01466210022031778

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cooper, C. R. (1977). Holistic evaluation of writing. In *Evaluating writing: Describing, measuring, judging.* (pp. 3–31). Urbana, IL: National Council of Teachers of English.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice*, *31*(3), 10–20.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, *52*(3), 177–193. https://doi.org/10.1037/h0044919

Csikszentmihalyi, M., & LeFevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, *56*(5), 815.

Elo, A. E. (1978). *The rating of chess players, past and present*. New York, NY: Arco Pub.

Florig, H. K., Morgan, M. G., Morgan, K. M., & Jenni, K. E. (2001). A deliberative method for ranking risks (I): Overview and test bed development. *Risk Analysis*, *21*(5), 913–921.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*(3), 193–202. https://doi.org/10.1037/0003-066X.39.3.193

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, *71*(3), 328–338.

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology*, *11*(1), 23–33. https://doi.org/10.1007/BF02278252

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management*, *48*(2), 251–268. https://doi.org/10.1002/hrm.20278

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, *6*(1), 48–60. https://doi.org/10.1177/1745691610393521

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research & Development in Education*, *27*(2), 73–82.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23.

Hazell, P. L., Tarren Sweeney, M., Vimpani, G. V., Keatinge, D., & Callan, K. (2002). Children with disruptive behaviours II: Clinical and community service needs. *Journal of Paediatrics and Child Health*, *38*(1), 32–40. https://doi.org/10.1046/j.1440-1754.2002.00715.x

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–19. https://doi.org/10.1007/BF03216919

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement*, *21*(3), 277–281. https://doi.org/10.1111/j.1745-3984.1984.tb01034.x

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing Performance: Designing, Scoring, and Validating Performance Tasks*. New York, NY: Guilford Press.

Jonesa, I., & Alcocka, L. (2013). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1–14.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics*, *89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*(1), 151–177.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 29–36). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1609835

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement: Interdisciplinary Research & Perspective*, *2*(3), 135–170. https://doi.org/10.1207/s15366359mea0203_1

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000/pdf

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, *22*(2), 135–155. https://doi.org/10.1007/s10798-011-9190-4

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., … Othman, A. R. (1998). Analytic Versus Holistic Scoring of Science Performance Tasks. *Applied Measurement in Education*, *11*(2), 121–137. https://doi.org/10.1207/s15324818ame1102_1

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. Presented at the Proceedings of the 21st annual international ACM …. Retrieved from http://dl.acm.org/citation.cfm?id=290965

Linacre, J. M. (1989). *Many-Facet Rasch Measurement* (2nd ed.). Chicago, IL: MESA Press.

Linacre, J. M. (1993). Generalizability theory and many-facet Rasch measurement. In *Annual Meeting of the American Educational Research Association*. Atlanta, GA.

Linacre, J. M. (1999). Understanding rasch measurement: Estimation methods for rasch measures. *Journal of Outcome Measurement*, *3*(4), 382–405.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. *Educational Testing Service R & D Connections*, *11*, 1–8.

Lloyd-Jones, R. (1977). Primary trait scoring. In *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Maydeu-Olivares, A., & Böckenholt, U. (2008). Modeling subjective health outcomes: top 10 reasons to use Thurstone's method. *Medical Care*, *46*(4), 346–348.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30.

McMahon, S., & Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 368–389. https://doi.org/10.1080/0969594X.2014.978839

Mi, F., & Yeung, D.-Y. (2015). Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9534

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

Moneta, G. B., & Csikszentmihalyi, M. (1996). The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality*, *64*(2), 275–310.

Morgan, K. M., DeKay, M. L., Fischbeck, P. S., & Morgan, M. G. (2001). A deliberative method for ranking risks (II): Evaluation of validity and agreement among risk managers. *Risk Analysis*, *21*(5), 923–937.

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation*, *8*(14), 1–8.

Mullis, I. V. S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement: Issues and Practice*, *3*(1), 16–18.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472–489. https://doi.org/10.1037/0033-295X.110.3.472

Oakes, M. E., & Slotterback, C. S. (2002). The good, the bad, and the ugly: Characteristics used by young, middle-aged, and older men and women, dieters and non-dieters to judge healthfulness of foods. *Appetite*, *38*(2), 91–97. https://doi.org/10.1006/appe.2001.0444

Odell, L., & Cooper, C. R. (1980). Procedures for evaluating writing: Assumptions and needed research. *College English*, *42*(1), 35–43. https://doi.org/10.2307/376031

Olson, J. M., Goffin, R. D., & Haynes, G. A. (2007). Relative versus absolute measures of explicit attitudes: Implications for predicting diverse attitude-relevant criteria. *Journal of*

Personality and Social Psychology, *93*(6), 907–926. https://doi.org/10.1037/0022-3514.93.6.907

Osborn Popp, S. E., Ryan, J. M., & Thompson, M. S. (2009). The critical role of anchor paper selection in writing assessment. *Applied Measurement in Education*, *22*(3), 255–271. https://doi.org/10.1080/08957340902984026

Pollitt, A. (2004). Let's stop marking exams. Presented at the IAEA Conference, Philadelphia.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, *55*, 72–75.

Rafoth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication*, *1*(4), 446–458.

Raman, K., & Joachims, T. (2014). Methods for Ordinal Peer Grading. *arXiv Preprint arXiv:1404.3656*. Retrieved from http://arxiv.org/abs/1404.3656

Raman, K., & Joachims, T. (2015). Bayesian Ordinal Peer Grading. Presented at the Proceedings of the Second ACM Conference on Learning @ Scale, New York, NY. Retrieved from http://www.cs.cornell.edu/~karthik/Publications/PDF/raman_joachims_14b.pdf

Rawson, K. A., & Middleton, E. L. (2009). Memory-based processing as a mechanism of automaticity in text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 353–370. https://doi.org/10.1037/a0014733

Robitzsch, A. (2017). *sirt: Supplementary item response theory models* [R package version 1.14-0].

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*(10), 1045–1063.

Sajjadi, M. S., Alamgir, M., & von Luxburg, U. (2015). Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. *arXiv Preprint arXiv:1506.00852*. Retrieved from http://arxiv.org/abs/1506.00852

Schamber, J. F., & Mahoney, S. L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *Journal of General Education*, *55*(2), 103–137. https://doi.org/10.1353/jge.2006.0025

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-tests. *Journal of Educational Measurement*, *51*(4), 400–418.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, *22*(2), 205–226. https://doi.org/10.1007/s10798-011-9194-0

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., & Wainwright, M. (2014). When is it better to compare than to score? *arXiv Preprint arXiv:1406.6618*. Retrieved from http://arxiv.org/abs/1406.6618

Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., & Ramchandran, K. (2013). A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*. Retrieved from http://lytics.stanford.edu/datadriveneducation/papers/shahetal.pdf

Sheppard, L. D., Goffin, R. D., Lewis, R. J., & Olson, J. (2011). The effect of target attractiveness and rating method on the accuracy of trait ratings. *Journal of Personnel Psychology*, *10*(1), 24–33. https://doi.org/10.1027/1866-5888/a000030

Shumate, S. R., Surles, J., Johnson, R. L., & Penny, J. (2007). The effects of the number of scale points and non-normality on the generalizability coefficient: A Monte Carlo study. *Applied Measurement in Education*, *20*(4), 357–376.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice*, *6*(3), 33–42. https://doi.org/10.1111/j.1745-3992.1987.tb00507.x

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., de Kruif, R. E. L., Reed, M., … White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, *59*(3), 492–506. https://doi.org/10.1177/00131649921970008

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology*, *38*(3), 368–389. https://doi.org/10.2307/1415006

Thurstone, L. L. (1959). *The Measurement of Values*. Chicago: University of Chicago Press.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, *9*(2), 1–10.

Wagner, S. H., & Goffin, R. D. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes*, *70*(2), 95–103. https://doi.org/10.1006/obhd.1997.2698

White, E. M. (1984). Holisticism. *College Composition and Communication*, *35*(4), 400. https://doi.org/10.2307/357792

White, E. M. (1985). *Teaching and Assessing Writing*. San Francisco, CA: Jossey-Bass.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco, CA: Jossey-Bass Publishers.

Wolcott, W., & Legg, S. M. (1998). *An overview of writing assessment: theory, research, and practice*. Urbana, IL: National Council of Teachers of English.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1–10. https://doi.org/10.1016/j.asw.2015.06.002

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.

Appendix A: Essay Prompt

Essay Prompt

More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends.

Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

Appendix B: Holistic Rubric

Holistic Rubric

Score Point 1: An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:

- Contains few or vague details.
- Is awkward and fragmented.
- May be difficult to read and understand.
- May show no awareness of audience.

Score Point 2: An under-developed response that may or may not take a position. Typical elements:

- Contains only general reasons with unelaborated and/or list-like details.
- Shows little or no evidence of organization.
- May be awkward and confused or simplistic.
- May show little awareness of audience.

Score Point 3: A minimally-developed response that may take a position, but with inadequate support and details. Typical elements:

- Has reasons with minimal elaboration and more general than specific details.
- Shows some organization.
- May be awkward in parts with few transitions.
- Shows some awareness of audience.

Score Point 4: A somewhat-developed response that takes a position and provides adequate support. Typical elements:

- Has adequately elaborated reasons with a mix of general and specific details.
- Shows satisfactory organization.
- May be somewhat fluent with some transitional language.
- Shows adequate awareness of audience.

Score Point 5: A developed response that takes a clear position and provides reasonably persuasive support. Typical elements:

- Has moderately well elaborated reasons with mostly specific details.
- Exhibits generally strong organization.
- May be moderately fluent with transitional language throughout.
- May show a consistent awareness of audience.

Score Point 6: A well-developed response that takes a clear and thoughtful position and provides persuasive support. Typical elements:

- Has fully elaborated reasons with specific details.
- Exhibits strong organization.
- Is fluent and uses sophisticated transitional language.
- May show a heightened awareness of audience.

Appendix C: Odd Scores Authentic Rubric Anchors

Anchor Essays in Authentic Rubric 1

Essay 1
Dear readers, I think that its good and bad to use the computer to much

Essay 2
Dear Newspaper, I think that the effects are okay as long as we get off the computers and go outside and see some friends and get some exercise. Computers let us not just talk to each other but it also lets us challenge each other on games without hurting each other it could even stop all ways all at once because we could challenge other countries in war games without killing real living people. We can check our taxes and stocks. It also makes it easier to find health insurance, car insurance, and house insurance. We can look up historical facts on the computer. We can look up how to stop snake venom from getting to your heart and how to make a how and some arrows to hurt with. You can find plumbers, technicians, oil companies, and lumber companies. You can find dates on the computer, too and find information about certain eople too.

Essay 3
Dear Newspaper, I strongly believe that technologies for kids and adults to use computers. I strongly believe that people have a bad affection people because you addicted, you spend less time with your family and friends. Finally, you spend less time excercising and doing great opportunities for your self. Clearly, i think that lettng students and adults use computers will take over their mind and make them addicted to the computer. This very bad because when your addicted to something, you can go back. This means you will have time for anything. For example, they micht forget what life is about and what was around for us to live in. Addiction is hard to get over , so don;t live your life in a box and play on nonsens live your life to the fullest. Over all, addiction in everything, including computers and a horrible . Without a doubt, spending too much time on computers will rott your brain and make you lazy. Meaning that the less you excersise, the less active you are. This problem is specially for teens. Teens have computer every week. cell phones, ipods, and many more. Alot of teens get cranky and the become node and neisty to their parents because they are testing to. As people dont excercise and move around alot, cause severe consequences like obesity and cancer. Obesity usually happens when you eating food, but it can also ahppen by not eating and not exercising. People that arent moving become lazy and have no desire to do anything which will an effect . Overall computers cause people to do the excercise and less active which will cause . Lastly, Computers make people spend less and less time with their family, and treat their computer as if the computers are . In everyday life , teens on their cell phones , their computers listening to , music . these electronics are causing. If this goes on alot of teens will loose the people they. Computers are a waste of time because they can make you. because they have affect on people who use computers and other electronics.

Appendix D: Even Scores Authentic Rubric Anchors

Anchor Essays in Authentic Rubric 2

Essay 1

Dear Newspaper, I think computers are a good thing, people use computers for buissness. Computers allow you to interact with people far away. They also allow you to do research. That is why I think computers are good. Computers can be used for many different things. They can be used for buissness. Some people work at home on computers. They can also be used to interact with people far away. You can interact with someone who you don't see on a daily basis. Computers can also be used for research. You can do research on a school project or on other buissnesses you interact with. Computers are good. Some people use computers for buissness. You can also interact with people far away. You can even do research on a computer that is why computers are good.

Essay 2

Do you spend all or most of your freetime sitting at your computer? People are spending too much time on computers. Obesity is a growing problem in today's society, people need to interact with eachother, and we need to get some fresh air. Computer are ruining these things for people. Unplug yourself from your computer. More and more people are becoming obese. This is do to spending too much time sitting down in front of computers. There is absolutely no exercise involved with being on your computer. People need to get outside and go jogging, play games or ride their bikes. Cutting back on computer time would cut back on the obesity problem in our society. Top scientists and researchers say that it is extremely important for people to interact with others face to face. To our whole life, we will have to interact and socialize with others. If you don't develop these skills, you won't be happy where you end up. Talking to people over the computer is nowhere near close to talking to someone face to face. Get out and get some fresh air. Go experience nature. Bring out on the beach, hiking in the woods or even just sitting on the swing in your backyard can be the most. Fresh air clear your mind. Time says that kids who spend time outside receive better grades than they would if they were inside on the computer. People need to releave stress so go outside. Click turn off your computers and get out of your chair. Go get some exercise, go interact with friendsand family, or even just go outside and get some fresh air while relaxing. People are wasting too much time on their computers and it's flushing our society down the toilet, so just turn off your computer.

Essay 3

The innovation of computers was a great leap into discovery. Unfortunately, the negative aspects of this magnificent invention detract from the novelty. With the computer came a plague of not exercising and a disinterest in, perhaps even an inabilty to enjoy nature. On top of this destruction of the mind, this monster is the cruel cause of many declining relationships for who has time for friends and family while in the middle of an internet game? As unique and intriguing as the computer might be, it is slowly destroying the mind. Increasing interest in a monitor tends to result in decreasing care for exercise as well as a decline in the amount of exercise gotten each day. While playing on the computer, the mind @MONTH1 be racing but the body sits unmoving, @CAPS1 activity is being done. Indeed, for some the addiction is awful enough that the person only moves to get food, use a restroom, go to bed or when a parent demands it. All of the day's food settles into the stomach and little to none of it is burned off. Meanwhile, in some of the more severe cases the person is so entranced that bedtimes are pushed later. Sleep is essential to the body and mind and going without can be damaging. Nowadays the computer detracts from the sleep and exercise that the body requires. Even nature suffered from the advent of computers. Fewer people went out to admire its beauty. Nature is a wonderful teacher about the circle of life and the natural way of the world. Oftentimes parallels can be drawn between the wild and the tamed. Attached as people are to computers, they cannot experience the world to the fullest. Some argue that flora and fauna can be found online, but seeing a picture is different, almost cheating, compared to being beside the plant, able to study it from all angles rather than in an unmoving form. The experience is not the same. Also, computers are indoors, so the true world is chielded from view. Even nowadays with laptops if one was to be brought outside, the user would be focusing on the screen, not what is around them. Computers have severely cut down on the admiration of nature. Even family and friends suffer when computers are around. Many people barely pay attention to conversation while focusing on the screen. They tend to onlypretend to listen half the time. Even when chatting with friends online, the conversation is disjointed because attention is probably focused on another part of the internet. Frequently, during conversations without a computer nearby, the mind is focused on something from earlier that day. Relationships are suffering, and computers are at fault. Computers are sucking in society, more of it each day. Much of what is done on computers, like the study of faraway places, can easily be found in books. Computers have resulted in a lack of exercise, less time with nature, strained relationships, and a generally unhealthy lifestyle. Through all of this, they have returned little to society. Why does the world continue to put up with computers and their negative effect on humanity?

Appendix E: Authentic Rubric Scale

Example of Authentic Rubics

Compare to the three example essays, how would you rank this essay?

Worst	As good as Essay 1	Between 1 & 2	As good as Essay 2	Between 2 & 3	As good as Essay 3	Best

Appendix F: Self-Report Questionnaire Items

End of Session Self-Report Items

1. How easy was it to evaluate the responses?

Very Hard	Hard	Fairly Hard	Moderate	Fairly Easy	Easy	Very Easy
1	2	3	4	5	6	7

2. How enjoyable was it to evaluate the responses?

Very Not Enjoyable	Not Enjoyable	Fairly Not Enjoyable	Moderate	Fairly Enjoyable	Enjoyable	Very Enjoyable
1	2	3	4	5	6	7

Appendix G: Demographic Items

Demographic Survey

For the purpose of your study, we would like to collect some basic demographic information. Your answers to theses questions will not affect your opportunity to participate in all phases of the current study.

1. How old are you (years)?

   ____

2. What is your Gender?

   Male Female

3. What is your highest education level?

   Did not finish high school
   High school diploma
   Some college
   Associate degree
   Bachelor degree
   Graduate degrees

4. How much teaching experience do you have (in years)?

   ____

5. Please indicate your experience in grading written assignment/test.
   Very Little Little Moderate High Very High

Appendix H: Sample Dataset

essay_id	rater1_domain1	rater2_domain1	Essay
14	3	3	My three detaileds for this news paper article is one state you opinion about the effects of computers. Seconde give detailed resons that will persuade of the local newspaper to agree with yor postition. This are my three ideas to the news paper article. To bigin my opinion about the computer effects are wast time. Many people wast time computers like fat people insted by insted by in the computer go and run or exercising. <truncated>
19	2	2	I aegre waf the evansmant ov tnachnolage. The evansmant ov tnachnolige is being to halp fined a kohar froi alnsas. Tnanchnolage waf ont ot we wod not go to the moon. Tnachnologe evans as we maech at. The people are in tnacholege to the frchr fror the good ov live. Famas invanyor ues tnacholage leki lena orde dvanse and his fling mashine <truncated>
49	3	3	Dear local newspaper, I am writing this letter to you to show on how people get effected from computers. It effects people very well and teaches people many things for just one technology. Computers effects many people for what it does. Computers have many advantages that people learn very quickly. Many people love computers. It is one of the most best inventions to everyone around the world. They teach positive things for people to learn. <truncated>

Vita

Wik Hung Pun graduated from Pennsylvania State University in 2008 with a Bachelor's degree in Psychology. After graduating from the undergraduate program, he studied in the Educational Psychology Program (measurement track) at the Pennsylvania State University. During this time, he taught multiple undergraduate level classes and served as statistical consultants for research programs.

In 2013, he earned a Master's degree in Educational Psychology at Pennsylvania State University in 2013, and he anticipates receiving a doctoral degree in May 2017. As of May 2017, he will be an assessment coordinator in the Learning Outcome Assessment department at Pennsylvania State University.