The Report Committee for IL PARK

Certifies that this is the approved version of the following report:

A Predictive Validity Study of AES system

APPROVED BY

SUPERVISING COMMITTEE

Supervisor: ______________________________________

Diane Schallert

____

Barbara Dodd

A Predictive Validity Study of AES systems

by

IL PARK, B.A., M.Ed.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

December 2010

Abstract

A Predictive Validity Study of AES systems

IL PARK, M.A.

The University of Texas at Austin, 2010

Supervisor: Diane Schallert

A predictive validity approach has been employed to find some implications to support evidences for Automated Essay Scoring (AES) systems. First, using R^2 values from multiple linear regression models, validity indices are compared first between multiple choice scores and essay scores across four AES systems. Secondly, R^2 values from models using only essay scores, the validity indices of four AES systems are hypothetically compared to see if how well AES systems could predict student outcome such as GPA.

Table of Contents

List of Tables

Chapter 1. Introduction

Writing assessment is not new, having been introduced around 2,100 years ago during the Chinese dynasty in selecting persons fit for higher positions(Page & Petersen, 1995). In modern days, writing assessment has been increasingly introduced in large-scale assessment for placement or selection purposes as well as in classroom assessment.

Writing assessment has come to be treated as a kind of performance assessment to the extent that it focuses more on the ability to demonstrate writing skill directly than on the ability to choose correct answers from among given choice options (Attali & Burstein, 2006; Bennett, 1993). For writing proficiency assessment, direct measures require an examinee to write an essay, typically on pre-selected topics, whereas indirect measures usually require an examinee to select an answer from among possible multiple-choice options (Barrett, 1994). Although direct measures are currently considered more desirable than indirect measures in writing proficiency assessment (Bennett, 1993), the increased use of direct writing assessments raises some issues, including cost, time, subjectivity, and scoring unfairness (Attali & Burstein, 2006; Bridgeman, Trapani, & Attali, 2009; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998). In addition, as a technique of performance assessment, direct writing assessment is exposed to psychometrics issues such as reliability and validity. For example, validity concerns require that the scores resulting from the assessment need to be interpreted in light of the evidence gathered to support inferences about the extent to which (a) scoring of responses is adequate, (b) the tasks are of sufficient domain coverage, and (c) the assessment result can be extrapolated to the target domain of interest (Chung & Baker, 2003; Kane, Crooks, & Cohen, 1999).

To address these complications in performance assessment, technological innovations have been useful in developing assessment methods. With innovations in task presentations, Automated Essay Scoring (AES) systems have been introduced to make possible performance assessment of many complex constructed-response items, replacing human scoring for the purposes of cost reduction, high objectivity, administrative efficiency, consistency, and impartiality (Burstein, et al., 1998; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001). Although AES systems have received some criticisms along the lines that they cannot judge, understand, or appreciate creative expressions, thinking flows, or exceptional and inspirational essays as well as humans can, they are considered more appropriate than traditional assessment (Attali & Burstein, 2006; Burstein, et al., 1998; Powers, et al., 2001). Especially if incorporated into a large-scale assessment, AES systems can perform well, not only for extraction of features to be built into scoring models, but also for presentation of reliable scores with reduced measurement errors.

With regard to measurement issues concerning AES systems, reliability and validity studies have been performed under the tacit precondition that human-generated scores are generally considered the gold standard, even though they may incorporate some random error (Chung & Baker, 2003). However, traditional methods of determining the reliability of AES are inappropriate, because random errors of measurement made by different human scorers are considered eliminated when scores are produced by AES systems (Keith, 2003; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). Though there may seem to be some trade-offs between systematic errors contained in AES scores and random measurement errors contained in human-generated scores, according to some recommendations, measurement accuracy related to AES systems could be more readily estimated through generalizability theory (Clauser, Kane, & Swanson, 2002; Yang, et al., 2002). Sources of threats to reliability regardless of scoring procedures, issues related to reliability unique to AES systems, and reliability from specific measurement contexts are issues primarily introduced in relation to reliability considerations (Cizek & Page, 2003).

Concerning validity issues, research has mainly focused on software validity, score validity, and assessment validity (Chung & Baker, 2003; Yang, et al., 2002). Whereas software validity deals with scoring algorithms incorporated in AES systems, score validity deals with the relationship between AES scores and human-generated scores. Assessment validity focuses on relationships between AES scores and external measurement criteria or tasks. Among these three types of validity studies, score validation studies have been the most frequent, with assessment validation rare, usually concerning the use of AES systems in real application with attention to validations for each AES system (Chung & Baker, 2003; Keith, 2003). In view of the scant research available about the assessment validity of AES scores, this report will concern predictive validation research, a form of assessment validity.

The purpose of this report is to investigate hypothetical predictive validity indices using four major AES systems to evaluate validation samples, in a real application context. Accordingly, it is appropriate for a review of literature and the research design to focus on a predictive study associated with four major AES systems, considering specifically measures and measurement issues for writing assessment, the different kinds of AES systems used, current validation studies, and predictive validity related to AES systems. Then, a discussion of a hypothetical research design, anticipated results, and implications will follow.

Chapter 2. Review and Critique of Existing Literature

Main focus of this chapter is on reviews and critiques about studies that have been conducted related to Automated Essay Scoring (AES) systems. Specifically, brief introductions are presented to compare measures that are available and some issues of validity for writing assessments, and introductions of AES systems in terms of development and usages that connect writing assessments with computer technologies, which employed in this study. Consequently, reviews on some research that have been performed employing AES systems will follow in view of validations, also some critiques will be made if applicable.

Direct measures vs. indirect measures for writing assessment

According to Barrett (1994), essay samples are considered direct measures of critical, higher order skills emphasized by the process approach that writing specialists have been making efforts to teach. However, direct measures require some decisions to be made to ensure score reliability, such as choice of scoring methods, (holistic or analytic), manner of score calculations, the design of the prompt for the essay, and the proper weighting of two sections when indirect measures are employed together with direct measures (Barrett, 1994). Direct measures have not been used as frequently as they might have been because analysis requires natural language processing, which requires a relatively higher amount of work (Powers, et al., 2001).

Nevertheless, as measures of writing skill, direct measures are usually favored over indirect measures because proponents believe that writing skills are better demonstrated through direct measures rather than indirect measures (Powers, et al., 2001). This is because the result is scores interpreted within the domain of writing, with the observed writing performances considered a sample (Kane, et al., 1999).

According to Burstein (2003), examples of large-scale assessment programs that employ "direct measures" and "indirect measures" for writing assessments include the Graduate Management Admissions Test (GMAT), Test of English as a foreign Language (TOEFL), Graduate Record Exam (GRE), Professional Assessment for Beginning Teachers (Praxis), Scholastic Assessment Test writing test, Advanced Placement (AP) exam, and the College Level Examination Program (CLEP) of English and Writing test. For some of these tests, computer based delivery methods have been introduced.

CONSTRUCTED RESPONSES FOR WRITING ASSESSMENT

A constructed response is broadly defined as any question requiring an examinee to generate an answer rather than select from a small set of options (Bennett, 1993; Powers, et al., 2001). In light of this definition, a constructed response seems to be consistent with the goal of direct measures in a writing assessment context because direct measures—that is, constructed response formats—focus on examinees' construction of their own authentic knowledge, whereas indirect measures do not. Examples of constructed responses include writing an essay, describing an experiment, or conducting an investigation (Bennett, 1993; Kane, et al., 1999).

Constructed response items use more authentic, real-life examples, and they are less vulnerable to specific test taking strategies or cheating. However, it is difficult to incorporate them into relatively large-scale assessments due to their high cost, demand on human resources, and delayed feedback associated with scoring by human scorers (Attali & Burstein, 2006; Bennett, 1993; Yang, et al., 2002). Across time and many testing occasions, these factors can bring about a certain degree of instability of scores due both to variability generated between human scorers and within any single scorer, introducing an additional source of measurement error (Yang, et al., 2002).

VALIDITY ISSUES FOR WRITING ASSESSMENTS

The most debated psychometric issues and the key points of distinction between direct measures and indirect measures are those related to reliability and validity. According to classical test theory, the reliability coefficient is defined as the correlation between strictly parallel tests, and it can never be determined but can be estimated for a given sample of subjects responding to a given sample set of test items (Crocker & Algina, 1986). The validity concerns the degree to which a test measures what it purports to measure (Lord & Novick, 1968), and is   described as the process by which a test developer or test user collects evidence to support the types of inferences that are supposed to be drawn (Crocker & Algina, 1986). Therefore, it is considered that validity must be judged in light of an ongoing process of developing a sound scientific basis for argument and gathering evidence that supports the proposed interpretation and actions based on the test scores (Clauser, et al., 2002; Yang, et al., 2002). Thus, an assessment procedure or a score from an assessment is neither valid nor invalid in itself; the inferences drawn from or the interpretation assigned to the scores are where the issue of validity takes place (Kane, et al., 1999).

Although both reliability and validity are important for assessing writing, in this report, focusing on the quality of writing, the inferences of scores, and the interpretation of scores generated by human vs. AES scorers, only validity issues are of concern.

Interestingly, in a predictive validity study comparing direct measures and indirect measures using classical test theory and item response theory, Barrett (1994) found that there were no differences in terms of validity coefficients indicated for direct measures or indirect measures in predicting a dependent variable. However, Powers, Burstein, Chodorow, Fowles, and Kukich (2000, 2001) found that a direct measure with constructed responses was more valid in assessing writing ability than an indirect measure, in spite of its being much more labor-intensive than using machine-scored, multiple-choice questions.

OVERVIEW OF AES (AUTOMATED ESSAY SCORING) SYSTEMS

In the early 1960s, Ellis B. Page, who was very interested in developing an essay scoring method using computers, developed Project Essay Grader (PEG) that was the very first AES system (Page, 1966, 2003). During early stages of development, scores from judges using PEG were not much different than those from other human scorers, and they were close enough to justify belief that computers had the potential to grade as reliably as English teachers (Page, 1966, 2003). Although interest in automated essay scoring methods using PEG never disappeared, studies of automated scoring system were not actively conducted during the 1970s and early 1980s. The advent of microcomputer systems after the mid-1980s and a variety of technological advances drove the re-examination of the potential for automated scoring methods (Page, 2003; Zenisky & Sireci, 2002).

As AES is defined as a computer system that assesses and generates scores for written prose (Dikli, 2006; Shermis & Burstein, 2003), there has been criticism that computers cannot replace humans because AES systems are only machines designed to do what humans command them to do, and AES systems cannot read, understand, or appreciate sentence structures, flow of thought, context, and content of sentences (Powers, Burstein, Chodorow, Fowles, & Kukich, 2000; Powers, et al., 2001). Further, computer infrastructure must be developed so that large numbers of students can use computers in preparation for writing assessment situations.

The goal of most AES systems is to emulate the best aspects of human scorers while minimizing the errors of human scorers, treating the essay prompts that are not easily dealt with by human scorers. In this respect, AES systems are supposed to be applied increasingly to large-scale assessment programs as well as classroom writing assessment. Though AES researchers are also interested in saving cost and time, and providing students and teachers with useful feedback (Dikli, 2006), AES systems are evaluated to be still weaker than human raters in scoring the content of essays and in assessing works in non-testing situations (Wild, Stahl, Stermsek, Penya, & Neumann, 2005). However, AES continues to draw attention from schools, universities, testing companies, researchers, and educators (Dikli, 2006).

AES systems include the PEG (Page, 1966) introduced earlier, e-rater (Burtein et al., 1998) e-rater v.2 (Attali & Burstein, 2006), Intellimetric^{TM} (Elliot, 2003), and Intelligent Essay Assessor (IEA, Landauer et al., 1998). Besides these automated *essay* scoring systems, there are similar systems that use computers to assess physicians' patient management skills, for an architect registration examination, and for dentistry assessments—systems using Criterion^{SM} , MY Access, Bayesian Essay Test Scoring System^{TM} (BETSY), and the Text Categorization Technique approach (Attali & Burstein, 2006; Burstein, 2003; Dikli, 2006; Dodd & Fitzpatrick, 1998; Srihari, et al., 2008; Yang, et al., 2002). For this report, only four systems, that is, PEG, e-rater, Intellimetric^{TM}, and Intelligent Essay Assessor (IEA) will be the focus, and they are introduced next.

INTRODUCTION OF FOUR AES SYSTEMS

PEG (Project Essay Grader). The PEG system developed by Ellis B. Page in 1964 was the first AES with the goal of predicting the scores that a number of competent human judges would give to a group of similar essays (Page, 2003). PEG was based on the assumption that the true quality of essays must be defined by human judges, although individual judges are not entirely reliable and may be biased. However, having more and more judges permits a better approximation to the true average rating of an essay.

Page proposed that some measurable features had to be extracted insomuch as constructs of interest in writing assessment are latent, not measurable directly (Yang, et al., 2002). Page started with student essays already graded by teachers and then experimented with a number of extractable textual features called "proxies," using multiple regression methods to determine which features best predicted the teachers' grades. PEG could then be used to score other essays using the same set of features.

Because some of the most predictive features were surface features such as word length, essay length in words, number of commas, number of prepositions, and number of uncommon words (Hearst, 2000; Page, 2003; Yang, et al., 2002), it was evident that indirect measures were being employed because of the computational difficulty of using direct measures. Therefore, critics such as Hearst (2000) stated that using indirect measures could leave PEG vulnerable to cheating and that indirect measures did not reflect important qualities of writing such as content, style, and organization, which could be the source of instructional feedback. PEG has more recently undergone transformation to include more direct measures of writing quality.

Although PEG provided only holistic scores initially, recently it has revised its data- collection and score-classification schemes to distinguish better students' writing abilities, and it now provides a kind of trait score for the organization or style of an essay, and for the purpose of instructional and diagnostic feedback (Yang, et al., 2002). In 2001, PEG offered a Web interface in order to assess writing abilities more effectively. The

Web PEG took only two minutes from submission to scoring, and it could assess three essays per second, with a correlation between human raters and PEG of .71, as compared to the correlation among human raters of 0.62. However, Yang et al. (2002) noted that the PEG website acknowledges that “PEG does not understand the content of your written product, but rather emulates how raters evaluate work that is similar to yours.”

PEG has been used on nationally normed tests that have substantial writing components, such as the Graduate Record Examination, Praxis, and NAEP (Page, 2003; Page & Petersen, 1995; Shermis, Koch, Page, Keith, & Harrington, 2002). In 1988, writing assessment using PEG was performed to obtain scores used for the National Assessment of Educational Progress (NAEP). PEG provided higher score reliability—correlation between human raters and PEG—than among human raters. In 1993, PEG was also used by Educational Testing Service (ETS) in scoring 1,314 Praxis exams that were the data source for a blind test to find a satisfactory degree of validity. In addition to these uses, PEG was also used for scoring in the Write America program, again providing higher score validity (.69) than human raters (.50).

IEA (Intelligent Essay Assessor). The Intelligent Essay Assessor (IEA) was developed by *Knowledge Analysis Technologies* in the late 1990s. It is a set of software tools for scoring the quality of the conceptual content of essays based on Latent Semantic Analysis (LSA).

LSA employs machine-learning methodology that develops a mathematical representation of the meaning relations among words and passages by means of statistical computations applied to a large corpus of text (Landauer, Laham, & Foltz, 2003; Srihari, et al., 2008). The technology is used to provide an accurate judgment of the semantic relatedness and similarity among documents or essays (Hearst, 2000; Srihari, et al., 2008; Yang, et al., 2002). LSA aims at going beneath an essay’s surface vocabulary to quantify its deeper semantic content (Hearst, 2000). The basic assumption of LSA is that the meaning of a passage is contained in its words, and that all its words contribute to a passage’s meaning. For example, this assumption means that even if one word of a passage is changed, its meaning may change, while conversely, two passages containing quite different words may have nearly the same meanings (Landauer, et al., 2003).

LSA serves as a way to determine the Euclidean distance between word meanings using cosine measures, and it uses empirical ratings from judges as the basis for determining the distances among words (Landauer, et al., 2003; Shermis, et al., 2002). LSA also permits the grader to set up a desired answer by having the software evaluate sections of text from a third source such as a textbook in setting the parameters for a desired outcome, so it can score even creative narratives equally well, even with few training sets (Landauer, et al., 2003; Shermis, et al., 2002).

According to Landauer et al. (2003) and Shermis et al. (2002), with regard to the relative prediction strengths of LSA and other measures, the LSA content measure was found to be the most significant predictor, far surpassing the indices of “style” and “mechanics” of five traits related to writing assessment such as content, organization, style, mechanics, and creativity. Although style and mechanics indices have strong predictive capacity on their own, when combined into a single index (the IEA total score), the content measure accounts for the most variance (e.g., 75% for all essays, 69% for standardized essays, and 79% for classroom essays) (Landauer, et al., 2003). IEA is known as the most well developed and widely used LSA based machine-scoring method (Srihari, et al., 2008). Besides using LSA, IEA also incorporates a number of other natural language processing methods to provide an overall approach to scoring essays and providing feedback (Landauer, et al., 2003).

According to Landauer et al. (2003), experiments performed using LSA measures to secure quality scores, using large sample sizes (total sample size 3,396: 2,263 in standardized GMAT test by ETS and 1,033 in classroom test at the University of Colorado), produced inter-rater reliability analyses comparing LSA scores to single reader scores with reliability coefficients around 0.86 on standardized tests and around 0.75 on classroom tests.

IntellimetricTM. Intellimetric^{TM} was developed with 10 years of experimental testing and released for commercial use in 1998 by Vantage Learning in affiliation with the College Board (Keith, 2003; Wang & Brown, 2007), the first AES system which was grounded on artificial intelligence technology. Using natural language processing (NLP) methods along with artificial intelligence (AI) and statistical technologies, the system first reads a pool of essays with known scores determined by expert raters. Then, it generates a unique scoring model based on a set of pre-scored responses without pre-specifying a set of features or rubrics (Elliot, 2003; Wang & Brown, 2007; Yang, et al., 2002).

According to Elliot (2003), Intellimetric^{TM} analyzes more than 300 semantic, syntactic, and discourse features in five major categories: “Focus and Unity,” “Development and Elaboration,” “Organization and Structure,” “Sentence Structure,” and “Mechanics and Conventions.” The category “Focus and Unity” is related to features pointing toward cohesiveness and consistency in purpose and main idea. “Development and Elaboration” is associated with features of text, looking at the breadth of content and the support for concepts advanced. “Organization and Structure” concerns features targeted at the logic of discourse, including transitional fluidity and relationships among parts of the response. “Sentence Structure” refers to features targeted at sentence complexity and variety. Finally, “Mechanics and Conventions” concerns features examining conformance to conventions of edited American English. Based on a holistic scoring algorithm, it assigns final scores to an essay writer based on these five major features. Intellimetric^{TM} is available for English, Spanish, Hebrew, Bahasa, Dutch, French, Portuguese, German, Italian, Arabic, and Japanese.

Elliot (2003) reported that Vantage Learning has conducted more than 140 studies in support of the use of Intellimetric^{TM} , including studies of admissions tests, entry tests, placement tests, literacy tests, medical performance tests, construct validity tests, and norm referenced tests, as well as studies of the relationship between Intellimetric^{TM} and multiple choice measures, between Intellimetric^{TM} and teacher judgments, and between Intellimetric^{TM} and other AES systems. In another of these studies, a rough report about the relationship between Intellimetric^{TM} and two other major AES was introduced (Elliot, 2003), showing that it provided somewhat greater scoring accuracy than the other two major AES examined. And, according to a study conducted by Wang and Brown (2007) to investigate the validity of Intellimetric^{TM} using 107 Hispanic participants from south Texas using WritePlacer *plus* test, a comparison of mean scores that were generated from Intellimetric^{TM} and human raters resulted in no significant differences.

E-rater. *E-rater,* based on Natural Language Processing (NLP) technology, was originally developed by ETS for the Analytical Writing Assessment section of the Graduate Management Assessment Test (GMAT) in late 1990s, and it has been in use since February, 1999 (Attali & Burstein, 2006; Bridgeman, et al., 2009; Burstein, 2003; Burstein, et al., 1998; Hearst, 2000; Powers, et al., 2000; Powers, et al., 2001).

As an application of computational methods to analyze characteristics of electronic files of text or speech, NLP technology is considered more relevant to the analysis of text-based applications. With the introduction of NLP technology, ETS could implement three independent modules for syntactic, discourse, and topical analysis through which *E-rater* can be used to extract key features related to a holistic scoring guide. For the syntactic module, a syntactic parser captures syntactic varieties by assembling phrases into trees based on sub-categorization information for verbs. For the discourse module, discourse identifiers are employed to determine the organization of ideas based on discourse classification schemas. For the topical analysis module, a vector-space model is used to capture the use of vocabulary for identification of the topic. Results from these three modules were found to correlate with essay scores provided by human raters (Burstein, 2003; Burstein, et al., 1998; Powers, et al., 2001; Yang, et al., 2002).

Counts from analysis of syntactic and discourse features and scores from analysis of topical features are computed and stored in vectors for model building and scoring. This procedure is called content vector analysis (CVA) which is a simpler form of LSA (Srihari, et al., 2008). E-rater is trained on approximately 300 hundred human rated essays for each question or prompt. All of the values resulted from CVA are subjected to linear regression to determine the optimal combination for building a model. Predictive features and their weightings are then provided by means of the regression method, and they are employed to assign a 6-point scale score using a step-wise linear regression. In total, 50-70 features are extractable, but in practice 8-12 features are retained for model building, and a different model is specified for each essay prompt (Burstein, 2003; Burstein, et al., 1998; Powers, et al., 2000; Srihari, et al., 2008).

E-rater scores tend to have more variability than human-generated scores because human scorers are likely to avoid providing extreme scores (Myford & Cline, 2002). In general, *E-rater* and one human rater are involved in making a score assignment. However, a third human rater is expected to make a settlement of wide differences, if the score discrepancy between *E-rater* and the human rater is greater than 1 point (Burstein, 2003; Powers, et al., 2001; Yang, et al., 2002).

According to Attali and Burstein (2006), E-rater V.2, which has been in use since 2006, differs from the previous version of e-rater and from other AES systems in contributing to its validity. One of the most important characteristics of e-rater V.2 is its use of a smaller set of meaningful and intuitive features, while the first version of e-rater used almost 60 features. Thus, the features are expected to be closely related to meaningful dimensions of writing, and the dimensions could be used for different scoring models. Consequently, single scoring model and standards can be used across all prompts of an assessment so that scoring procedures can be successfully applied on data from a couple of essays of the same assessment. Attali and Burstein (2006) also stressed that E-rater V.2 could possibly control the construction of scores in terms of meaning and external evidence associated with the performance of the different dimensions. Also, E-rater V.2 has been embedded in on-line writing websites that allow users to practice this (Lee, Gentile, & Kantor, 2008).

Interestingly, Powers et al. (2001) found that experts might be successful in misleading the system of e-rater V.1 by tricking the computer program (for example, by cutting and pasting the same sentences or passages). Bridgeman, et al (2009) conducted a study to investigate the validity and fairness of the E-rater V.2 by comparing the differences in mean scores generated by human raters and E-rater across ethnic groups using 11th grade English Test, Test of English as a Foreign Language (TOEFL), and Graduate Record Exam (GRE). There were some inconsistencies for certain ethnic groups (Chinese students), even though human and e-rater scores for most subgroups were comparable.

COMPARISON OF AES SYSTEMS

A comparison of the four AES systems is shown in Table 1.

Table 1 A comparison of the four AES systems

	PEG	IEA	Intellimetric^{TM}	E-rater
Developer (year)	Ellis B. Page (1964)	Knowledge Analysis Technology (late 1990's)	Vantage Learning (1996)	ETS (1998, v1) (2006, v2)
Embedded Technology	NLP	LSA & NLP	NLP	NLP
Training sample of essays	Needed	Not needed	Needed	Needed
Statistical Techniques	Regression (Simultaneous)	Regression (hierarchical / sequential)	Artificial Intelligence	Multiple Linear Regression (stepwise)
Scoring Procedures	Holistic procedure		Holistic procedure	Holistic procedure
Modules (units) for extracting features	Surface feature		Semantic Syntactic Discourse	Syntactic Discourse Topical
On-line Embedded?				Yes

Note. NLP = Natural Language Processing; LSA = Latent Semantic Analysis.

CURRENT VALIDATION STUDIES FOR AES SYSTEMS

Three validity issues are associated with AES: “software validity,” “score validity,” and “assessment validity” (Chung & Baker, 2003; Keith, 2003; Yang, et al., 2002).

Software validity refers to the validation of software incorporated or embedded in the AES system in order for the system to execute the technical requirements of its investigation. Such validation concerns the computer system technology and related program language for the score processing algorithms.

Score validity concerns the correspondence between scores generated by human raters and by AES systems to find evidence whether the system provides valid scores as compared to human-generated scores. Inter-rater agreement correlation coefficients are used to express comparability. Although high values of these indexes are usually reported regardless of the AES system used, this approach is deemed independent of AES's application context.

Assessment validity focuses on evidence that using the system supports identified assessment goals, a judgment about the adequacy of the intended use of the assessment results. Assessment validity will be the focus of the proposed study, in which administration of an assessment to a sample from a population of interest through a standardized procedure will take place, with evaluation of the assessment results with respect to its intended use. Focusing on this sort of validity is expected to provide realistic information about the usefulness of AES scores in an application context. Among the goals of this report is investigation of the practical relationships between AES scores and external criteria to see whether AES systems closely reflect the intended use of the scores in real world situations.

A study by Keith (2003) has provided clues about comparable validities among AES systems as well as detailed evidence concerning AES score validity. Keith conducted a comprehensive study of the relative validity of AES systems, examining validity indices for four AES systems: PEG, Intellimetric^{TM}, Intelligent Essay Assessor, and E-rater. Examining correlations between AES systems and human judges, Keith found that values for PEG ranged from .48 to .86, those for Intellimetric^{TM} ranged from .68 to .88, those for IEA ranged from .65 to .90, and those for E-rater ranged from .69 to .85. Examining correlations between AES systems and other measures of writing, Keith found correlations between Intellimetric^{TM} essay scores and multiple-choice test scores ranging from .55 to .69, and those between Intellimetric^{TM} essay scores and teacher ratings of writing skills ranging from .46 to .76. Examining the correlation between IEA scores from undergraduates' heart anatomy and function essay and a short answer test, Keith found a value of .76. Finally, Keith found that the E-rater system showed relatively lower correlation coefficients than did the Intellimetric^{TM} and IEA systems, showing, for example, values ranging from .09 to .27 using external criteria, .27 for self-reported grades in a writing course, and .24 for undergraduate writing samples.

Chapter3. Needed Research

As described earlier, validity studies involving AES scores have been performed for three processes that include software validation, score validation, and assessment validation. Although much evidence has been gathered that supports the validity of scores produced by AES systems, rarely have AES scores been evaluated in situations of practical application. Undoubtedly, the correlation between AES scores and human generated scores is important in judging the reliability as well as the validity of the growing use of AES systems. However, as Chung and Baker (2003) claimed, whereas each AES system provides relatively high validity, this finding is not enough to serve as evidence for the validation of AES scores. They also insisted that criterion-referenced validity studies are needed in the context of practical application of AES systems.

Keith (2003) also suggested that relative validation research among AES systems still remains to be done. He implied that a cross-program blind test between AES systems using both screening and calibration samples should be performed, so that the results could provide empirical information about the relative validity of and possible improvements for each AES system.

Therefore, in the proposed research here, a research method is proposed for investigating predictive validation of AES systems concerned with assessment validation. The method is designed to determine not only the relative contributions of direct and indirect measures incorporated into AES systems but also the relative values of indices of predictive validation.

Due to the large volume of essays that must be scored on time for major assessments, AES systems based on computer technologies are now being used to score writing samples. As recently as 1993, over 9 million essays were scored by human raters at ETS (Page & Petersen, 1995). The growing use of AES systems in practical applications has resulted from their cost effectiveness, impartiality, stability, and other advantages over the use of human scorers.

However, according to Keith (2003), it is unclear whether AES systems have validity in general, whether AES systems measure essay skills, or whether all AES systems are equally valid. And score validation studies have been conducted far more frequently than studies concerned with assessment validity in a situation of practical application. As a part of an assessment validity study, Keith (2003) found evidence of validity for each of four AES systems: PEG, Intellimetric^{TM}, IEA, and E-rater. In addition, he addressed the question of the relative validity of AES systems through a blind test using a validation sample. Keith (2003) and Chung and Baker (2003) addressed the predictive validation of AES scores using a validation sample to see if AES systems are successful in an application context.

The following study is proposed in order to address two principal research questions:

1. Do essay scores contribute more than multiple-choice scores to the prediction of writing assessment scores when AES systems are used?

2. Are predictive validity coefficients consistent among AES systems?

To answer the first research question, a hypothetical predictive validity index will be computed for the correlation between AES scores consisting of essay and multiple-choice scores (as predictor variables) and subsequent English course grade (as dependent variable) across four AES systems. To answer the second question, virtual comparisons will be made between predictive validity indices for AES scores and subsequent English course grade across four AES systems.

METHOD

Sample Description

Two samples of 1,000 undergraduate freshmen at the University of Texas at Austin for the 2009-2010 academic year will be selected. The sample is divided into two groups which are presumed to have randomly assigned 1,000 subjects respectively. For cross-validation purpose, one group is used as a calibration sample, and the other as a validation sample (Lord & Novick, 1968; Pedhazur, 1997).

Instrument score

English language and composition scores provided by the College Board Advanced Placement Program (AP) will be the source of scores for this study. The AP examination is taken by high school students in an effort to receive college course credit or advanced placement, or both. The English language and composition examination contains a multiple-choice section and a free-response section. The scores from the two sections are combined to compute a composite score ranging from 1 to 5. The interpretations for assigned scores, as provided by the College Board, are 5, extremely qualified; 4, well qualified; 3, qualified; 2, possibly qualified; and 1, no recommendation (Dodd, Fitzpatrick, & De Ayala, 2002).

Variable descriptions

Four components of the AP English language and composition test scores, which consist of three essay scores (E1, E2, and E3) obtained using four AES systems and a multiple-choice score (MC) will be used as independent variables. The grade from a subsequent college English course (E316K) for undergraduates will be used as the dependent variable.

AES systems employed

Four AES systems will be used to generate the four AP test scores: Project Essay Grader, Intellimetric^{TM} , Intelligent Essay Assessor (IEA), and E-rater.

Research Design

Table 2 Research Design

AES systems		Independent Variables	No	Data	
				Sample 1(N_1=1,000) (Calibration Sample)	Sample 2(N_2=1,000) (Validation Sample)
A	E-rater	E1	1	GPA	GPA
		E2	2		
		E3	3		
		MC	4		
B	Project Essay Grader (PEG)	E1	1	GPA	GPA
		E2	2		
		E3	3		
		MC	4		
C	Intellimetric^{TM}	E1	1	GPA	GPA
		E2	2		
		E3	3		
		MC	4		
D	Intelligent Essay Assessor (IEA)	E1	1	GPA	GPA
		E2	2		
		E3	3		
		MC	4		

Note. E1=Essay Prompt1; E2=Essay Prompt2; E3=Essay Prompt3; MC=Multiple Choice Score; GPA= Grade of E316K.

Models

The multiple linear regression equations for the calibration and validation samples are introduced separately below, distinguishing between the full model and restricted model, and also applied across the four AES systems. The full model consists of a dependent variable (GPA_{09}) and four independent variables that are scores from three essay prompts and a multiple choice response; the restricted model is made up of the same dependent variable and three independent variables that are three essay prompt scores only. For the two models, the multiple linear regression equations for the screening sample and the calibration sample can be expressed as follows:

[Full models for the two samples]

$$GPA_{cal} = \beta_{0i} + \beta_{ij} \cdot X_{ij} \text{ , } \textit{(i=1,2,3,4; j=1,2,3,4)} \qquad (1)$$

$$GPA_{val} = \beta_{0i} + \beta_{ij} \cdot X_{ij} \text{ , } \textit{(i=1,2,3,4; j=1,2,3,4)} \qquad (2)$$

[Restricted models for the two samples]

$$GPA_{cal} = \beta_{0i} + \beta_{ij} \cdot X_{ij} \text{ , } \textit{(i=1,2,3,4; j=1,2,3)} \qquad (3)$$

$$GPA_{val} = \beta_{0i} + \beta_{ij} \cdot X_{ij} \text{ , } \textit{(i=1,2,3,4; j=1,2,3)} \qquad (4)$$

where, ***cal*** refers to the calibration sample, ***val*** is the validation sample, ***i*** designates the AES system used (1= E-rater; 2 = PEG; 3 = Intellimetric^{TM}, and 4 = IEA), ***j*** means the *j*th independent variable (0=intercept; 1=Essay prompt1; 2=Essay prompt2; 3=Essay prompt3; 4=Multiple Choice score), and β_{0i} is the intercept given equation.

ANALYSIS

Predictive validity refers to the degree to which test scores predict criterion measurements that will be made at some point in the future (Crocker & Algina, 1986). Where two or more independent variables are concerned, the predictive validity coefficient between a set of predictors (or independent variables) and a criterion is equal to the correlation between the criterion and the particular linear combination of the predictors that minimizes the squared error of prediction (Lord & Novick, 1968).

In this context, the multiple correlation coefficient *R* is considered a measure of predictive validity in the specific sense of the minimization of the squared error of prediction. For example, if the multiple correlation *R* from a regression model using as a dependent variable college GPA and as independent variables SAT score and high school GPA is 0.40, then college GPA has a some degree of predictive validity with respect to SAT and high school GPA (Crocker & Algina, 1986; Lord & Novick, 1968).

The multiple correlation coefficient *R* and its squared value of R^2 will be adopted for abating validity indices. The multiple correlation coefficient *R* indicates the degree of the linear relationship between "predicted" values of the English course E316K grade (estimated grade) and "observed" values of English course 316K Grade (observed grade) from a regression using a set of AES component scores as independent variables, and a GPA as the dependent variable. Therefore, R is needed as a predictive validity coefficient to measure the linear relationship between multiple independent variables (E1, E2, E3, and MC) and a dependent variable (grade). However, instead of estimating the population multiple correlation, a cross-validation method is used to see if a regression from a sample (that is, the screening sample) performs in another sample (namely, the calibration sample) from the same population.

In doing so, specifically, a regression is performed for the screening sample first. The regression equation from the screening sample will be applied to the calibration sample. Then, using the same predictors as were used for the screening sample, the estimated value of dependent variable (grade) will be generated for each subject in the calibration sample. Thus, a linear correlation between the observed grade and the predicted grade is calculated. This correlation is considered as a cross-validity coefficient (R').

Also, the multiple correlation of R for the screening sample and the cross-validation coefficient (R') for the calibration sample will be computed across the four AES systems so that the relative predictive value of scores produced under each system can be determined. Values of R^2 and the squared cross-validation coefficient $(R')^2$ for each AES system will also be calculated, where R^2 indicates the amount of variance in the grade that is accounted for by the set of multiple independent variables in the screening sample and the squared cross-validation coefficient [$(R')^2$] is the amount of variance in the GPA$_{09}$ which is explained by the set of the same independent variables in the calibration sample. In that sense, R^2 and $(R')^2$ are expected to show the strength of linear relationship between the English course GPA and the AES-produced set of component scores for the screening and calibration sample respectively.

In addition, the values of R^2 and $(R')^2$ will be used to test the difference in hypothetical significance between two proposed models, such as the full or restricted model, separately for the screening and the calibration sample, using the generalized F test. For example, R^2 will be used to compare the full model having E1, E2, E3, and MC scores as independent variables and English course GPA as dependent variable with a restricted model having only E1, E2, and E3 scores as independent variables and the same dependent variable.

Formulae for the computation of R^2 (and R is simply the square root of R^2), F-test statistics for hypothetical significance testing for R, and the F statistic for model comparison using R^2 are introduced below (Lord & Novick, 1968, p.268; Pedhazur, 1997, pp.105-109, 983):

$$R^2 = 1 - \frac{|C|}{|C_x|} \qquad (5)$$

where, $|C|$ is the determinant of the correlation matrix of all the variables, namely, the independent variables as well as dependent variable, and $|C_x|$ is the determinant of the correlation matrix of the independent variables.

$$F = \frac{\left(R^2\right)/K}{\left(1 - R^2\right)/(N - K\ \ - 1)} \qquad (6)$$

$$F = \frac{\left(R^2_{Grade.1234} - R^2_{Grade.123}\right)\Big/(K_{full} - K_{res})}{\left(1 - R^2_{Grade.1234}\right)\Big/(N - K_{full} - 1)} \tag{7}$$

where, ***N*** means sample size, ***K*** is the number of independent variables, and ***full*** or ***res*** specifies the Full model or the Restricted model used respectively. As may be noticed, a general F-test can also be used for the calibration sample using $(R\ )^2$ instead of R^2 in the formulae.

If the difference between the R^2 of the screening sample and the $(R\ )^2$ of the calibration sample is small, the regression equation obtained from the screening sample can probably be applied for future predictions, presuming the conditions stay unchanged. A regression equation from the combined samples (namely, combining the screening and calibration sample together) is also expected to be used for stability of future prediction.

ANTICIPATED RESULTS

First, to address the first research question, comparisons of the full and restricted models across four AES systems will be performed to see if direct measures (writing efficiency) are better predictors of course GPA than indirect measures (multiple-choice options). Specifically, across the four AES systems, a hypothetical F-test for significance of difference between the two models will be performed with one degree of freedom, because the restricted model has only 1 fewer predictor variable than the full model.

Based on the findings of previous studies (Bennett, 1993; Powers, et al., 2000; Powers, et al., 2001) that found evidence that direct measures are more favorable than indirect measures, it is expected that the restricted model, having only direct measures as independent variables, will be a better predictor. That is, it is expected that the F-statistic for the relationship between the full model (using all of the independent variables) and restricted model (using only the direct measures) will not be statistically significant. If this expectation is not supported, the results would be consistent with the findings of Barrett (1994). Also, through the cross-validation, the values of R^2 and $(R')^2$ across four AES systems could be compared to see if there is any substantial difference for the regression from the screening sample to be used for future prediction in finding usefulness of direct measures rather than indirect measures under the same conditions. Also, if there is no difference found between R^2 and $(R')^2$, the more stable regression could be obtained from the combined sample.

Second, to address the second research question, hypothetical tests will be performed to investigate whether predictive validity coefficients are different from 0 and statistically significant. The multiple linear correlation between course grade and three direct measures (from only the restricted models) within each AES system will be examined. Even lacking clear criteria for judging whether the validation coefficient is high or low, according to Page (2003) and Keith (2003), it is also expected that the validation coefficient for the relationship between the dependent variable and the set of independent variables within each AES system is expected to be significantly different from 0. If this expectation is not supported, the applicability of the scores from each AES system should be reconsidered. Likewise, in terms of predictive validation, the values of R^2 and $(R')^2$ only from the restricted models could be compared to see whether the predictive validity could be consistent across four AES systems. Likewise, if there is no difference found between R^2 and $(R')^2$ from the restricted models, the more stable regression for four AES systems could be obtained from the combined sample.

Expected results are shown in Table 3, in the form of answers to Research Questions 1 and 2 across the four AES systems, for a 2x2x4 display (sample x model x AES system).

Table 3 Summary of anticipated results across four AES systems

AES systems	Research Questions	Model	Validity Index	
			Sample 1(N_1=1,000) (Calibration Sample)	Sample 2(N_2=1,000) (Validation Sample)
E-rater	Q1	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
	Q2	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
Project Essay Grader (PEG)	Q1	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
	Q2	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
Intellimetric^{TM}	Q1	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
	Q2	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
Intelligent Essay Assessor (IEA)	Q1	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}
	Q2	Full	R_{Full} & R^2_{Full}	R'_{Full} & R'^2_{Full}
		Restricted	R_{RES} & R^2_{RES}	R'_{RES} & R'^2_{RES}

Note. R_{Full}=Multiple correlation coefficient from full model; R_{RES}=Multiple correlation coefficient from restricted model; R_{COM}=Multiple correlation coefficient from combined model.

Chapter4. Conclusion

Understanding that existing validation studies using AES systems have mainly been conducted focused on score validity, software validity, and assessment validity, this report has been focused on one form of assessment validity, the predictive validation context. Specifically, it is performed hypothetically to obtain evidences by employing writing assessment scores generated from AES systems to support whether AES could also be useful in predicting student outcomes such as GPA.

The first research question was raised to investigate which is more predictive between direct measures and indirect measures when using AES systems. The second question addressed   implications to see if the AES systems have sound predictive validity so as to be practically used by consumers such as teachers, schools, testing companies, and project investigators. To answer the questions, regression analysis would be conducted using hypothetical data, and cross-validation processes are then performed using R^2 and R'^2 given models across four AES systems.

Based on the anticipated results of this hypothetical report, several points may be discussed regarding the first question about whether AES systems could be employed as useful tools to find better models in predicting students' outcomes using direct and indirect measures. As introduced earlier, though direct measures are usually favored over indirect measures, direct measures require some decisions to be made such as choice of scoring methods, (holistic or analytic), manner of score calculations, the design of the prompt for the essay, and the proper weighting of two sections when indirect measures are employed together with direct measures (Barrett, 1994; Powers, et al., 2001). In this context, if AES systems are found to be useful in predicting students' outcomes using direct measures, time and energy concerning these decision-making processes could be considerably saved.

With regard to the second question, some results could be supportive of what AES system works better than others in terms of predicting students' outcomes. So far, not many research have been done to compare predictive validation across AES systems used in this report except Keith (2003). Though, this hypothetical study cannot provide specific ranges of validity coefficients, the results could be compared with the findings done by Keith (2003). He found that correlations between AES systems and other measures of writing, the correlations of Intellimetric^{TM} essay scores ranged from .55 to .69 (with multiple-choice scores), from .46 to .76 (with teacher ratings of writing skills), the correlation of IEA scores (from undergraduates' heart anatomy and function essay) and a short answer test was .76. The E-rater system showed relatively lower correlation values ranging from .09 to .27 using external criteria, .27 for self-reported grades in a writing course, and .24 for undergraduate writing samples.

If the predictive validity coefficients of AES systems are demonstrated, those coefficients may be helpful to support the use of AES in other fields which employ qualitative research methods, such as interviews or open-ended questionnaires. Among other uses, evidence of predictive validity will offer the promise of using AES scores as instructional feedback even as they are used as a standard to find a sound AES system, lightening the load for teachers who wish to administer writing assessments.

In addition, if some features of AES are to be extracted for practical use, perhaps current score modeling techniques in AES could also be applied to fields beyond those that simply perform word-by-word translation, tape recording, and data coding.

In terms of predictive validation, even though the samples are presumed to be randomly drawn, as the correlation coefficient could be affected by the variance of sample, the validity coefficient calculated from the University of Texas students might need to be corrected for attenuation due to the homogeneity in score distribution. Therefore, the numeric information cannot generalize the result with regard to its use. Also, as this research will be conducted hypothetically, specific predictive validity coefficients will not generally be useful for comparing the quality or usefulness of the four AES systems used, making it difficult to venture a recommendation of which system is better than another. Though a better one among the four AES systems could be chosen, cost effectiveness issue still remains.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-31.

Barrett, M. J. (1994). *Predictive validity of direct and indirect methods of writing assessment: A comparison using classical test theory and item response theory.*, University of Texas at Austin, Austin, TX.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. W. Bennett, B. C. (Ed.), *Construction Versus Choice in Cognitive Measurement* (pp. 1-28). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bridgeman, B., Trapani, C., & Attali, Y. (2009). *Considering fairness and validity in evaluating automated scoring*.

Burstein, J. (2003). The e-rater scoring engine: automated essay scoring with natural language processing. In J. S. Burstein, M. D. (Ed.), *Automated Essay Scoring – A cross-disciplinary perspective* (pp. 113-121). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). *Computer analysis of essays*. Paper presented at the NCME Symposium on Automated Scoring, April 1998. from http://www.ets.org/research/erater.html

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated essay scoring of constructed responses. In J. S. Burstein, M. D. (Ed.), *Automated Essay Scoring – A cross-disciplinary perspective* (pp. 23-40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., & Page, E. B. (2003). The concept of reliability in the context of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 125-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement In Education, 15*(4), 413-432.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*: Holt, Reihart and Winston, Inc.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 2006-2012.

Dodd, B. G., & Fitzpatrick, S. J. (1998). *Alternatives for scoring computer-based tests.* Paper presented at the ETS colloquium, Computer-based testing: Building the foundation for future assessments.

Dodd, B. G., Fitzpatrick, S. J., & De Ayala, R. J. (2002). *An investigation of the validity of AP grades of 3 and a comparison of AP and non-AP student groups.* (No. 2002-9). New York, NY: College Entrance Exam Board.

Elliot, S. (2003). IntelliMetricTM : From here to validity. In J. S. Burstein, M. D. (Ed.), *Automated Essay Scoring – A cross-disciplinary perspective* (pp. 71-86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hearst, M. A. (2000). The debate on automated essay grading. *Ieee Intelligent Systems & Their Applications, 15*(5), 22-27.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In J. S. Burstein, M. D. (Ed.), *Automated Essay Scoring – A cross-disciplinary perspective* (pp. 147-168). Hillsdale, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with Intelligent Essay AssessorTM. In J. S. Burstein, M. D. (Ed.), *Automated Essay Scoring – A cross-disciplinary perspective* (pp. 87-112). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lee, Y., Gentile, C., & Kantor, R. (2008). Analytic Scoring of TOEFL?CBT Essays: Scores From Humans and E-rater? *Educational Testing Service. TOEFL Research Report RR?1. Retrieved May, 12*, 2009.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Myford, C. M., & Cline, F. (2002). *Looking for patterns in disagreements : A facets analysis of human rater's and e-rater'sTM scores on essays written for the Graduate Management Admission Test (GMAT).* Paper presented at the the annual meeting of the American Educational Research Association, April 1-5, 2002.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238-243.

Page, E. B. (2003). Project Essay Grade : PEG. In J. S. Burstein, M. D. (Ed.), *Automated Essay Scoring – A cross-disciplinary perspective* (pp. 43-54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Page, E. B., & Petersen, N. (1995). The computer moves into essay grading: updating the ancient test. *Phi Delta Kappan*, *76*(7), 561-565.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York, NY: Holt, Reinhart, & Winston.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). *Comparing the validity of automated and human essay scoring* (No. GRE No.98-08aR). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-Rater : Challenging the validity of automated essay scoring* (No. GRE Board Professional Report No. 98-08bP).

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring : a cross-disciplinary perspective*. Mahwah, N.J.: Laurence Earlbaum Associates.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(1), 5-18.

Srihari, S., Collins, J., Srihari, R., Srinivasan, H., Shetty, S., & Brutt-Griffler, J. (2008). Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence*, *172*(2-3), 300-324.

Wang, J., & Brown, M. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology, Learning, and Assessment*, *6*(2).

Wild, F., Stahl, C., Stermsek, G., Penya, Y., & Neumann, G. (2005). Factors Influencing Effectiveness in Automated Essay Scoring with LSA. In C. K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.), *Artificial Intelligence in Education - Supporting Learning through Intelligent and Socially Informed Technology* (Vol. 125, pp. 947-949).

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement In Education*, *15*(4), 391-412.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement In Education*, *15*(4), 337-362.

Vita

IL PARK was born on January 29, 1969 in Busan, Repulic of Korea to Korean parents whose names are Jaehong Park (father) and Jeongja Cho (mother). He also graduated from Seoul National University with degrees of B.A.(1997) in Education and M.Ed.(2000) in Educational Method major. He also worked at the Electrical Technology Research Institute (ETRI) as a HRD administrative staff for two years from 2000 to 2002.

Permanent address: ilpark703@gmail.com

This report was typed by IL PARK.