교육과정평가연구
The Journal of Curriculum and Evaluation
2014, Vol. 17, No. 3, pp. 247~269

Agreement rates between automated essay scoring systems and human raters: Meta-analysis[1)]

In-Soo Shin(Assistant Prof. Jeon-Ju Univ.)*

≪ ABSTRACT ≫

Automated essay scoring (AES) is defined as the scoring of written prose using computer technology. The objective of this meta-analysis is to consider the claim that machine scoring of writing test responses agrees with human raters as much as humans agree with other humans. The effect size is the agreement rate between AES and human scoring estimated using a random effects model. The exact agreement rate between AES and human scoring is 52%, compared with an exact agreement rate of 54% between humans. The adjacent agreement rate between AES and human scoring is 93%, compared to an adjacent agreement rate of 94% between humans. This meta-analysis shows that the agreement rate between AES and human raters is very comparable. This study also compares the subgroup analysis of agreement rates using study characteristic variables such as publication status, AES type, essay type, exam type, human expertise, country, and school level. Implications of the results and potential future research are discussed in the conclusion.

Key words : automatic essay scoring, agreement rate, meta-analysis, effect size

1) This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government(NRF-2012S1A5A8023737).
* 교신저자: s9065031@jj.ac.kr

Ⅰ. Introduction

Since the 1960s, machine scoring of essay tests has developed dramatically and is still developing. Automated essay scoring (AES) is defined as the scoring of written prose using computer technology (Shermis & Barrera, 2002). The importance of essay-based exams is that they are representative of real world tasks. As a result, this form of examination has become a standardized test for large-scale assessment (Enright & Quinlan, 2010).

However, there are difficulties surrounding essay scoring. Kelly (2001) states that the increasing demand of essay-based exams creates a heavy burden on test scoring. Cost and effort are the largest obstacles to adopting essay-based exams. AES is mainly used to overcome such difficulties, including time, cost, and reliability (Bereiter, 2003). Most importantly, an AES system should give the same score as would be given by humans.

Compared to human scoring, an AES lacks transparency, and does not give evidence of quality in the same way a human would score, evaluate, and understand a written essay (Kelly, 2006; Bennett, 2006). AES systems are currently used in several applications. Since February 1999, the Educational Testing Service (ETS) has used *e-rater* as one of its two initial raters for the Graduate Management Admission Test (GMAT) writing assessments, and, in this capacity, it has scored over 1 million essays (Chodorow & Burstein, 2004). The ETS uses *e-rater* to score essays for numerous educational institutions, covering secondary and tertiary education.

Skepticism and criticism have accompanied AES over the years, often related to the fact that a machine cannot understand written text (Page & Petersen, 1995), a lack of human interaction (Hamp-Lyons, 2001), vulnerability to cheating (Rudner & Gagne, 2001), and the need for a large corpus of sample text to train the system (Chung & O'Neil, 1997). Despite its weaknesses, AES continues to attract the attention of public schools, universities, testing companies, researchers, and educators (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998). There have been many validation studies to date. Partly in response to critiques of the AES, there is a growing body of literature on the attempts to validate the meaning and uses of AES (Yang, Buckendahl, Jusziewicz, & Bhola, 2002).

The findings of these validation studies are contradictory, reporting mixed results. Bennett (2006) showed that the automated scoring of essay responses did not agree with the scores awarded by human raters in a National Assessment of Educational Progress (NAEP) study.

The results of McCurry (2010) suggest that essay marking software cannot score an open writing task as reliably as human raters can in the Australian Scaling Test (AST) writing test. Nichols (2005) also indicated a stronger relationship between two human raters than between the Intelligent Essay Assessor (IEA) and a human rater.

Despite these negative findings, many positive studies, report a high level of agreements between AES and human scoring. For current AES systems, these comparisons have shown impressively high levels of agreement with human scoring (Attali, 2004; Burstein & Chodorow, 1999; Elliot, 2003; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003; Vantage Learning, 2000a, 2000b, 2001a, 2001b, 2002, 2003a, 2003b), often comparable to those found between two human raters. There are many empirical studies for the agreement rate between human raters and an AES. However, no meta-analysis study has been conducted for the agreement rates.

The objective of this meta-analysis is to consider the claim that machine scoring of essay-based exams agrees with those of human raters as much as humans agree with other humans, as investigated by McCurry (2010). The aim is not to replace human raters. The goal of an AES system is to simulate a human expert rater's grading process, and a system is usable only if it can perform the grading as accurately as an expert human rater can (Chen, Liu, Lee, & Chang, 2010). Exact agreement means that the scores given by the AES and the human raters match perfectly, while an adjacent agreement means that the scores differ by at most one point. The exact and adjacent agreement rates are better validation indices for this aim. Most studies have used correlation coefficients as an index to investigate the relationship between human raters and an AES. However, correlation is inadequate in investigating the accuracy of AES grading because correlation does not represent the exact degree of agreement (e.g., adacent agreement rate). Correlation shows the relationship between the human rater and the AES, whereas the exact agreement rate pinpoints the exact degree of consistency between the humans and the AES. To investigate the agreement rate between AES and human raters systematically, meta-analyses are an appropriate methodology (Borenstein, Hedges, Higgins, & Rothstein, 2009). Meta-analytic procedures refer to a set of statistical techniques used to systematically review and synthesize independent studies within a specific area of research. Glass (1976) first proposed such methods and coined the term "meta-analysis."

There have been no previous meta-analysis studies that have investigated the agreement rate between AES and human scoring. The meta-analysis studies of Bergstrom (1992), Kim (1999), Mead and Drasgow (1993), Wang, Jiao, Young, Brooks, and Olson (2007), and Wang

et al (2008) examined the test mode effect on tests that measure general aptitude, ability, and achievement. The test mode effect is the discrepancy between performance in Computer-based tests (CBT) and paper-based tests (PBT), even when the tests are identical. In addition to these meta-analyses, Mazzeo and Harvey (1988) conducted a review of the literature on the comparison of scores from automated and conventional educational and psychological tests. While these studies investigated the test mode effect, the present article is the first to consider a different meta-analysis, the objective being, systematic examination of how AES systems compare with human raters. A further aim of the article is to discuss future research in AES and how it could be developed (Borensteinetal, 2009). The questions to be addressed as follows:

(ⅰ) What is the exact agreement rate between AES and human scoring?
(ⅱ) What is the adjacent agreement rate between AES and human scoring?
(ⅲ) What is the difference between the exact agreement rate and the adjacent agreement rate between AES and human scoring, and how does this compare with that between human raters?
(ⅳ) What is the moderator effect of categorical variables such as reporting characteristics (publication status, country), test characteristics (AES type, essay type, exam type), and subject characteristics (scorer expertise, student school level)?

Ⅱ. Materials and methods

Although meta-analysis is not a primary research study, it shares common research procedures in terms of the formulation of a problem, collection of data (primary studies in this case), coding of data, analysis, and interpretation (Cooper, 2010).

1. Literature search and inclusion criteria

A. Literature search

Searching began with existing meta-analyses and literature reviews. A thorough search was conducted on the Education Resources Information Center (ERIC), PsychInfo, Web of

Science, Google Scholar, and Digital Dissertations, to target empirical articles that fit the inclusion criteria.

The following keywords were used: automatic scoring, automated scoring, *e-rater*, essay, writing, and agreement. The literature was selected based on the abstracts. We also employed the "snowball method" and reviewed the references in the selected articles for additional works. Furthermore, we gathered review articles and theoretical overviews and checked their references. In all, we found 15 studies for analysis based on the inclusion and exclusion criteria given in the following subsection.

B. Inclusion criteria

While automatic scoring can be applied to essay writing, speaking, and other forms of examination, this meta-analysis considers only automatic scoring for essay writing because it is the most developed form of examination (Shermis & Burstein, 2003). Data were obtained to compute the effect size, which is the agreement rate between AES systems and human raters estimated using random effects model. The exact and adjacent agreement rates are better validation indices, though most studies use correlation as an index to investigate the relationship between human raters and the AES. Correlation is inadequate in investigating the relationship between human raters and the AES; however, exact agreement rates show the exact degree of consistency between humans and the AES systems. Exact agreement means that the scores given by the AES system and the human raters matched perfectly, and the adjacent agreement rate means that the scores differed by at most one point. This meta-analysis includes published, peer-reviewed journal articles and unpublished articles such as dissertations and conference papers to minimize the publication bias. Only articles written in English were considered because most AES systems are developed in the US and the UK (McCurry, 2010). Although additional non-empirical literature and literature reviews were selected as sources of relevant research, this literature was not included in the analysis because they have no information on calculating effect sizes.

2. Coding of studies

Data were extracted from studies meeting the inclusion criteria. Whenever studies reported multiple effects, only those that met the review criteria were included. Study characteristics were coded to reflect potential moderating variables for the agreement rate

between AES systems and human raters (Table 1). These characteristics included reporting characteristics (publication status, country), test characteristics (AES type, essay type, exam type), and subject characteristics (scorer expertise, student school level). Two coders independently coded each study. A coding manual was developed to help maintain reliability. The manual included information on the effect size calculation, study characteristics, and report characteristics. If there were discrepancies between the two coders, they tried to reach a consensus. However, when the discrepancies remained unresolved, the two coders discussed the differences, and a third, independent coder made a resolution. Finally, differences between the two coders were unanimously resolved prior to data entry and analysis

3. Computation of effect sizes

For this meta-analysis, the common metric used was effect size of proportion or percentage agreement. All effect sizes were calculated with the aid of the comprehensive meta-analysis (CMA) with inputs confirmed by two coders. The analysis was also carried out using CMA2.0 software to estimate the mean effect size (Borenstein et al, 2009).

According to Card (2011) the agreement rate calculation and analysis formulae are as follows:

Agreement rate (Event rate) = Events/Total

LogitEventRate = log (p / (1 - p))

LogitEventSE = Sqr (1 / (p * Total) + 1 / ((1 - p) * Total))

EventRate = (e ^ LogitEventRate) / (e ^ LogitEventRate + 1)

Here, e = 2.718281828, the value of the exponential function at 1.

For example, if the total number of papers is 63 and the scorers agree on 56 of them, we have the following:

p = 56 / 63 = 0.889

LogitEventRate = log(0.889 / (1 - 0.889)) = 2.081

LogitEventSE = Sqr (1 / (0.889 * 63) + 1 / ((1 - 0.889) * 63)) = 0.401

EventRate = (2.718281828 ^ 2.081) / (2.718281828 ^ 2.081 + 1) = 0.889

(This example calculation was performed in CMA 2.0)

Analyses were performed on the logit value, weighted by the standard error of

"LogitEventSE" as described here (Card, 2011). According to Card, the proportion works well as an effect size in many situations, but is problematic when it strays too far from 0.5. For this reason, it is useful to apply a logit to p prior to a meta-analysis. The event rate can be recovered from the logit value by applying the inverse of the logit function, as seen in the example shown here. Effect sizes are reported as 1 when agreement exists between the AES and human scoring on all papers. If there is agreement on none of the papers, the agreement rate is 0.

4. Combining effect sizes across studies

Once an effect size was calculated for each study, effects testing the same hypothesis were averaged. Weighted analyses, developed by Hedges and Olkin (1985), and the fixed effects model, the random effects model, and subgroup analysis, could be used for analysis.

In the weighted procedure, more weight was given to effect sizes with larger samples, on the assumption that the larger samples more closely approximated actual effects (Cooper, 2010; Hedges & Olkin, 1985). These weighted combined effect sizes were tested for statistical significance by calculating the 95% confidence intervals (Cooper, 2010).

The fixed effects model allows only for generalization to the study sample, whereas the random effects model allows for generalization to a larger population (Sirin, 2005). The fixed effects model assumes that primary studies have a common effect size. On the other hand, the random effects model attempts to estimate the distribution of mean effect size based on the assumption that each primary study may have different populations. For this study, the random effects model was used for the main effect and for subgroup analysis, because the homogeneity test was statistically significant (Borenstein et al, 2009).

In a meta-analysis, reviewers assume that every primary study is independent, but, there may be interdependence within the study if a study has multiple effect sizes. One could, in this case, use the same sample repeatedly, but this would violate the independence assumption. Alternatively, one might consider choosing an effect size among the multiple effect sizes within a study in an attempt to avoid violating the independence assumption, but this would result in loss of information. Cooper (2010) proposed a "shifting unit of analysis", where one estimates a total effect size, taking the study itself as a unit; however, if effect sizes for subgroups are estimated instead, each effect size is a unit. For this meta-analysis, the unit of analysis is a study for the estimation of the total effect size, whereas the unit of

analysis is an effect size for the subgroup analysis, according to Cooper's method. This strategy is a compromise that allows studies to retain maximum information value, while minimizing the violation of the independence assumption.

Ⅲ. Results

1. Description of effects

The method described previously provided 98 effect sizes from 10 primary studies for the exact agreement rate between AES and human scoring, 66 effect sizes from 13 primary studies for the adjacent agreement rate between AES and human scoring, 39 effect sizes from 9 primary studies for the exact agreement rate between human raters, and 34 effect sizes from 12 primary studies for the adjacent agreement rate between human raters. Because the studies exhibit multiple outcomes, it is important to be careful about the interdependence of these outcomes.

2. Overall analysis

In the homogeneity test (Table 2), the effect sizes of the primary studies were heterogeneous. Therefore, we measured the overall effect size using the random effects model and compared the effect sizes using the characteristics of each study (e.g., publication status, AES type, essay type, exam type, scorer expertise, country, school level.).

The random effects model (Cooper, 2010) is stated as follows:

$$\widehat{\sigma_\theta^2} = s^2(g) - (1/k)\sum_{i-1}^{k}\nu_i, \text{ where } s^2(g) = \sum_{i-1}^{k}\frac{(g_i - \overline{g})^2}{(k-1)}.$$

The exact agreement rate between the AES and human scoring was 52% with 95% confidence intervals at 50% and 54% (Table 3). The exact agreement rate between human raters was 54% with 95% confidence intervals at 50% and 57%.

The adjacent agreement rate between AES and human scoring was 93%, with 95% confidence intervals at 91% and 95%. The adjacent agreement rate between human raters

was 94% with 95% confidence intervals at 91% and 96%. These results show that there is a high agreement rate between AES and human scoring, compared to the agreement rate between human raters.

3. Subgroup analysis

This study performed a subgroup analysis using a random effects model, because each subgroup was heterogeneous based on homogeneity tests. Subgroup analysis was conducted to identify the source of variability and moderators, which affect the extent to which subgroups differ. We briefly discuss the subgroup analysis of each of the categorical variables.

A. Subgroup analysis for reporting characteristics

In meta-analysis, publication bias is an important issue for valid study results. The publication bias means that studies having statistically significant results have more possibilities of being published than non-significant studies do. In Table 4, the effect size of unpublished studies is higher than that of published studies, but it is not statistically significant. The effect size for the US is higher than that for the UK and Australia (Table 4). The different effect sizes between countries may be a result of the different examination types among countries. The UK exam is more open-ended, while Australia's exam has a larger scales (i.e., a 10-point scale). The use of the open-ended exam type is more difficult to score consistently compared to the closed-ended exam type. The scoring scale in the US is a 6 point-scale, but Austrailia has a 10-point scale. As the number of scale values increases, the consistency of agreement rate decreases.

B. Subgroup analysis for test characteristics

The agreement rate can be influenced by test characteristics such as different AES systems, essay types, and exam types.

In order of decreasing exact agreement rate, when comparing AES systems and human scoring, the AES systems used were IntelliMetric, *e-rater*, IEA, and Bayes. For the adjacent agreement rates (again, AES compared with human scoring), the order was IntelliMetric, IEA, *e-rater*, and Bayes. There was no statistically significant difference between the AES

systems.

In order of decreasing exact agreement rate for AES, compared with human scoring, the essay types were issue, narrative, expository, persuasive, and argument. The order was the same for adjacent agreement rates. In human-to-human comparisons, the order for both the exact agreement rate and the adjacent agreement rate was argument, issue, narrative, expository, and persuasive. The essay type may demand different content, structure, wording, level of knowledge, and writing style. The number of effect sizes is small.

In order of decreasing exact agreement rate when comparing AES to human scoring, the exam types were GRE, statewide exams, TOEFL, and GMAT. In all other situations (adjacent rates in AES-to-human comparisons, and exact and adjacent rates in human-to-human comparisons) the order was TOEFL, GRE, statewide exams, and GMAT (Table 5).

C. Subgroup analysis for subject characteristics

The agreement rate can be influenced by subject characteristics such as a different level of scoring expertise of scorer, and test taker's schooling level.

The effect size for AES-to-expert comparisons is higher than that for AES-to-non expert comparisons, and expert's scoring was more consistent. For the effect of schooling level on AES-to-human comparisons the school levels were undergraduate had the highest effect followed by graduate, K-12, and non-native (ESL) in decreasing order. For the adjacent agreement rate in human-to-human comparisons, the order was graduate, undergraduate, K-12, and ESL (Table 6).

4. Meta-regression by publication year

The agreement rate can be influenced by publication year, which makes it possible to investigate whether the present study has a better agreement rate or not.

The slope of meta-regression by publication year is positive, and statistically significant (Table 7). Thus, the general trend is that the agreement rate increases over time.

Ⅳ. Discussion and Conclusion

The meta-analysis in this article shows that there is good agreement between AES and human scoring. The exact agreement between AES and human scoring is 52%, compared with 54% for exact agreement between human raters. For the adjacent agreement rates, these figures become 93% and 94% respectively.

In the subgroup analysis for reporting characteristics, the effect size of unpublished studies is higher than that of published studies, but it is not statistically significant. The effect size for the US is higher than that for the UK and Australia. Traditionally, there is a transatlantic difference in the philosophy and mechanism of assessment, with the US placing more emphasis on "objective" multiple-response tests, and the UK on essay-type responses (Hutchison, 2007). However, according to Hutchison, multiple-choice questions now constitute a substantial proportion of the UK General Certificate of Secondary Education, conversely, in the US, constructed responses are increasingly used as part of large-scale assessments. The different effect sizes between countries may result from the differences in exam type in this meta-analysis, such as those discussed previously. The UK exam in this meta-analysis is composed of more open-ended essays than the US exam is. The AST program began as the Australian Scholastic Aptitude Test (ASAT) in the early 1970s, with the aim of bringing the assessments of different colleges in the Australian Capital Territory to a common scale (McCurry, 2010). The Australian Scaling Test of Writing test (ASTW) is an atypical writing test. It is an assessment of verbal reasoning and writing ability in which candidates are asked to respond in an argumentative mode to a broad range of stimulus material on a social or political issue (McCurry, 2010). ASTW is scored on a 10-point scale. Because the ASTW follows a different format from many other exams, caution should be exercised when analyzing it, and the exam type and the schooling level should be considered. Further studies should investigate the difference of the country as a moderator. However, we should acknowledge that most developments of AES systems are carried out in the US.

In the subgroup analysis for test characteristics, the order of AES type, by decreasing effect size was IntelliMetric, *e-rater*, IEA, and Bayes. Even though the scoring mechanisms used by these systems differed, they were not statistically different. The order of essay type, by decreasing effect size, was issue, narrative, expository, persuasive, and argument.

The type of essay is a potential source of the difference in scoring of essays. The essay type may demand different content, structure, wording, level of knowledge, and writing style. The number of effect sizes is small, so further primary studies are needed. The exam types, in order of decreasing exact agreement rates, were GRE, statewide exams, TOEFL, and GMAT. For the adjacent agreement rate, the order was TOEFL, GRE, statewide exams, and GMAT. Like essay types, exam types can also affect results. The relationship between logic and complexity of the essay and the AES system should be examined in further detail. The difference of agreement rates could depend on the level of rigidity in essay scoring criteria. Which should be investigated in future studies, as well as the roles of countries, and essay and exam types.

In the subgroup analysis for subject characteristics, the exact agreement rate between AES and scoring by experts is higher than that between AES and scoring by novice raters. This is similar to the result of Nichols (2005), which indicates a stronger relationship between the IEA and experts than between novice raters and experts. The school levels, in order of decreasing effect size, were undergraduate, graduate, and K-12.

As Burstein and Chodorow (1999) suggest that developers of AES systems should further investigate how to reliably score non-native test takers' written essays, the agreement rate for non-native (ESL) essay scoring is lower than that of other K-12 scores. This is a major issue to developing the AES, as teaching English to non-natives is a very important issue.

The fact that the slope of the meta-regression by publication year is positive (and statistically significant) suggests that AES development will continue to result in improvements in the agreement rate. While the present meta-analysis contributes to the evidence based approach for the development of AES in essay grading, the goal of future work should be to make an even larger and more sophisticated range of features considered in essay scoring available. As AES technology is still developing (Shermis & Burstein, 2003) the search for better machine scoring is ongoing as investigators continue to move forward in their drive to increase the accuracy and effectiveness of AES systems.

The present study has several limitations:

First, there are several validation indices, such as correlations, kappas, standardized mean differences, and agreement rates. However, this study used only agreement rates because they show the exact amount of consistency, compared to other indices.

Second, only papers written in English were considered, yet agreement rates may depend on the language that the tests are written in. The interpretation should be cautious because of the potential biases for studies written in English, which means that English studies have

a high effect tendency (Higgins & Green, 2009). Nowadays, many countries and language groups are now developing their own automatic scoring systems (McCurry, 2010; Nichols, 2004; Rudner, & Gagne, 2001; Shermis, & Burstein, 2003), so further studies are needed to investigate how language affects the reliability of AES.

Third, in order to ensure that results are not biased, more testing needs to be carried out, independent of developers of AES systems, because many of the current validity studies have been performed by the developers themselves.

Fourth, this study has limitations on the investigation of individual characteristics of the AES system. Future studies should be conducted on the agreement rate, based on the characteristics, logic, and complexity of the AES.

Fifth, this study includes only AES systems from countries such as US, and the UK. Therefore, it can not be generalized to other countries (e.g.,Korea), because the inclusion criteria consider only English articles. Most AES systems are developed by the US and the UK and further studies are required for other countries and language.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion.* Paper presented at the National Council on Measurement in Education (NCME), SanDiego, CA.

*Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *Journal of Technology, Learning, and Assessment (JTLA), 4*(3), 1-29.

Bennett, R. E. (2006). *Technology and writing assessment: lessons learned from the US national assessment of educational progress.* Annual Conference of the International Association for Educational Assessment, Singapore, IAEA.

Bereiter, C. (2003). Foreword. In Mark D. Shermis and Jill C. Burstein. (Eds.), *Automated essay scoring: A cross disciplinary approach* (vii-ix). Mahwah, NJ: Lawrence Erlbaum Associates.

Bergstrom, B. (1992). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-anaylsis.* West Sussex, UK:Wiley.

*Burstein, J., & Chodorow, M. (1999). *Automatedessay scoring for nonnative English speakers.* Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). *Computer analysis of essays.* Proceedings of the NCME Symposium on Automated Scoring, Montreal, Canada.

*Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (2001). *Enriching automated essay scoring using discourse marking.* ETS Reports-Evaluative.

Card, N. A. (2011). *Applied meta-analysis for social science research.* The Guilford Press.

Chen, Y., Liu, C., Lee, C., & Chang, T. (2010). An unsupervised automated essay scoring system. *Intelligent Sytems, IEEE. 25*(5), 61-67.

*Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays.* Research report. RR-04-04. ETS.

Chung, G. K., & O'Neil, H. F. Jr. (1997). *Methodological approaches to online scoring of essays.* ERIC Document Reproduction Service No. ED 418 101.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL, 212*(2), 259-279.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach (4th ed.).* Thousand Oaks, CA: Sage.

*Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-34.

*Elliot, S. (2003). IntelliMetric: From here to validity. In Mark D. Shermis and Jill C. Burstein. (Eds.). *Automated essay scoring: A cross disciplinary approach,* 71-86. Mahwah, NJ: Lawrence Erlbaum Associates.

*Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing, 27,* 317-334.

Hamp-Lyons, L. (2001). Fourth generation writing assessment. In T. Silva and P. K. Matsuda. (Eds.), *On second language writing,* 117-125. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando FL: Academic Press.

Higgins, J., & Green, S. (2009) *Cochrane handbook for systematic review of interventions.* John Wiley & Sons, Ltd.

Hutchison, D. (2007). An evaluation of computerized essay marking for national curriculum assessment in the UK for 11-year-olds. *British Journal of Educational Technology, 38*(6), 977-989.

*Kelly, P. A. (2001). *Computerizing scoring of essays for analytical writing assessments: Evaluating score validity.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Kelly, P. A. (2006). Review of the book automated essay scoring: A cross-disciplinary perspective. *Applied Psychological Measurement, 30*(1), 66-68.

Kim, J. (1999). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein. (Eds.),

Automated essay scoring: A cross-disciplinary perspective, 87-112. Hillsdale, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). *How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans.* Proceedings of the 19th Annual Conference of the Cognitive Science Society, 412-417. Mahwah, NJ: Erlbaum.

*Larkey, L. S. (1998). *Automatic essay grading using text categorization techniques.* Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval (90-95), Melbourne, Australia.

*Lonsdale, D., & Strong-Krause, D. (2003). *Automated rating of ESL essays.*

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

*McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing, 15,* 118-129.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 9,* 287-304.

Nichols, P. D. (2004). *Evidence for the interpretation and use of scores from an automated essay scorer.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.

*Nichols, P. D. (2005). *Evidence for the interpretation and use of scores from an automated essay scorer.* Research Report. RR-05-02. Pearson Educational Measurement.

Page, E. B. (2003). Project essay grade: PEG. In M. Shermis & J. Burstein. (Eds.), *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

*Raminenia, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012a). *Evaluation of the e-rater scoring engine for the GRE issue and argument prompts.* Research Report, RR-12-06. ETS

*Raminenia, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012b). *Evaluation of the e-rater scoring engine for the TOEFL independent and integrated prompts.* Research Report, RR-12-02. ETS

Rudner, L., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer* (ERIC Digest number ED 458 290).

Shermis, M., & Barrera, F. (2002). *Exit assessments: Evaluating writing ability through automated essay scoring* (ERIC document reproduction service No ED 464 950).

Shermis, M., & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Vantage Learning. (2000a). *A study of expert scoring and IntelliMetric scoring accuracy for dimensional scoring of grade 11 student writing responses* (RB-397). Newtown, PA: Vantage Learning.

Vantage Learning. (2000b). *A true score study of IntelliMetric accuracy for holistic and dimensional scoring of college entry-level writing program* (RB-407). Newtown, PA: Vantage Learning.

Vantage Learning. (2001a). *About IntelliMetric* (PB-540). Newtown, PA: Vantage Learning.

Vantage Learning. (2001b). *Applying IntelliMetric technology to the scoring of 3rd and 8th grade standardized writing assessments* (RB-524). Newtown, PA: Vantage Learning.

Vantage Learning. (2002). *A study of expert scoring, standard human scoring and IntelliMetric scoring accuracy for statewide eighth grade writing responses* (RB-726). Newtown, PA: Vantage Learning.

Vantage Learning. (2003a). *Assessing the accuracy of IntelliMetric for scoring a district-wide writing assessment* (RB-806). Newtown, PA: Vantage Learning.

Vantage Learning. (2003b). *How does IntelliMetric score essay responses?* (RB-929). Newtown, PA: Vantage Learning.

Vantage Learning. (2003c). *A true score study of 11th grade student writing responses using IntelliMetric version 9.0* (RB-786). Newtown, PA: Vantage Learning.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67,* 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5-24.

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

• *논문접수 : 2014-08-16/ 수정본접수 : 2014-09-29/ 게재승인 : 2014-10-13*

〈Table 1〉 Characteristics of studies included in the analysis

Author(Year)	Exact	Adjacent	H-H Exact	H-H adjacent	AES	Exam
Enright(2010)	0.570	0.980	0.560	0.970	e-rater	Toefle
Chodorow(2004)	0.493	0.943	0.560	0.960	e-rater	Toefle
Larkey(1998)	0.485	0.913	0.560	0.950	Bayes etc.	N.A.
Attali(2006)	0.510	-	0.507		e-rater	Toefle /GMAT
Raminenia(2012)a	0.558	0.967	0.585	0.975	e-rater	GRE
Raminenia(2012)b	0.512	0.944	0.600	0.975	e-rater	Toefle
Hutchison(2007)	0.407	0.871	0.425	0.858	e-rater	U.K. exam
Dikli(2006)	0.607	0.967	0.547	0.957	IntelliMetric	State exam
Nichols(2005)	0.509	0.946	0.491	0.932	IEA	State exam
Elliot(2003)	0.617	0.983	-	-	IntelliMetric	N.A.
McCurry(2010)	-	0.824	-	0.889	MSW	Australia AST
Lonsdale(2003)	-	0.670	-	0.660	NLP	ESL
Burstein(1999)	-	0.923	-	-	e-rater	Non-native
Kelly(2001)	-	0.880	-	0.960	e-rater	GRE
Burstein(2001)	-	0.935	-	-	e-rater	GMAT

〈Table 2〉 Results of the homogeneity test

Category	AES/Human	N	Q	*p*-value	-95%CI	ES	+95%CI	I^2
Exact Agreement	AES/Human	10	294.9	< .05	0.519	0.515	0.523	96.9
	Human/Human	9	349.3	< .05	0.534	0.539	0.544	97.7
Adjacent Agreement	AES/Human	13	794.5	< .05	0.940	0.942	0.944	98.5
	Human/Human	12	841.9	< .05	0.936	0.939	0.942	98.7

N = number of studies; *Q*=homogeneity statistic; p-value for Q statistics; CI=confidence interval; *ES*=effect size.

〈Table 3〉 The overall result of meta-analysis using a random effects model

Category	AES/Human	N	-95%CI	ES	+95%CI
Exact Agreement	AES/Human	10	0.497	0.521	0.544
	Human/Human	9	0.501	0.537	0.573
Adjacent Agreement	AES/Human	13	0.905	0.929	0.948
	Human/Human	12	0.907	0.939	0.961

N = number of studies; CI = confidence interval; *ES*=effect size

〈Table 4〉 Subgroup analysis for reporting characteristics

Category	Sub-group	Sub-outcomes	*k*	-95%CI	ES	+95%CI
AES/Human EXACT	Publication	No	54	0.497	0.519	0.540
		Yes	44	0.491	0.508	0.525
Human/Human EXACT		No	25	0.502	0.525	0.547
		Yes	14	0.453	0.489	0.525
AES/Human ADJACENT		No	61	0.933	0.945	0.955
		Yes	6	0.781	0.894	0.952
Human/Human ADJACENT		No	29	0.931	0.947	0.959
		Yes	5	0.792	0.901	0.956
AES/Human EXACT	Country	UK	3	0.384	0.407	0.429
		USA	95	0.503	0.518	0.532
Human/Human EXACT		UK	3	0.370	0.424	0.480
		USA	36	0.503	0.521	0.538
AES/Human ADJACENT		Australia	2	0.747	0.824	0.881
		UK	3	0.854	0.871	0.885
		USA	62	0.934	0.946	0.956
Human/Human ADJACENT		UK	3	0.824	0.859	0.889
		USA	30	0.932	0.948	0.960

K = number of effect sizes; CI = confidence interval; *ES*=effect size

〈Table 5〉 Subgroup analysis for test characteristics

Category	Sub-group	Sub-outcomes	*k*	-95%CI	ES	+95%CI
AES/Human EXACT	AES type	e-rater	40	0.500	0.514	0.528
		IEA	15	0.473	0.513	0.553
		IntelliMetric	9	0.536	0.613	0.684
		Bayes	5	0.399	0.492	0.580
AES/Human ADJACENT		e-rater	20	0.928	0.949	0.965
		IEA	15	0.938	0.953	0.964
		IntelliMetric	9	0.947	0.962	0.973
		Bayes	5	0.849	0.926	0.965
AES/Human EXACT	ESSAY type	Argument	24	0.462	0.484	0.507
		Expository	9	0.458	0.516	0.574
		Issue	7	0.513	0.547	0.580
		Narrative	3	0.400	0.523	0.644
		Persuasive	3	0.472	0.494	0.515

Category	Sub-group	Sub-outcomes	*k*	-95%CI	ES	+95%CI
Human/Human EXACT		Argument	4	0.496	0.547	0.597
		Expository	9	0.457	0.490	0.523
		Issue	2	0.458	0.546	0.633
		Narrative	3	0.494	0.518	0.542
		Persuasive	3	0.463	0.485	0.506
AES/Human ADJACENT		Argument	21	0.883	0.914	0.938
		Expository	9	0.930	0.952	0.967
		Issue	4	0.940	0.974	0.989
		Narrative	3	0.923	0.963	0.982
		Persuasive	3	0.917	0.945	0.964
Human/Human ADJACENT		Argument	4	0.937	0.957	0.970
		Expository	9	0.915	0.934	0.948
		Issue	2	0.973	0.980	0.985
		Narrative	3	0.938	0.953	0.965
		Persuasive	3	0.905	0.927	0.943
AES/Human EXACT	Exam type	State wide	18	0.489	0.528	0.567
		Toefle	8	0.512	0.520	0.528
		GRE	6	0.527	0.559	0.589
		GMAT	8	0.466	0.484	0.501
Human/Human EXACT		State wide	18	0.481	0.501	0.521
		Toefle	3	0.588	0.599	0.611
		GRE	2	0.570	0.585	0.600
		GMAT	2	0.470	0.495	0.520
AES/Human ADJACENT		State wide	18	0.942	0.955	0.965
		Toefle	4	0.913	0.967	0.988
		GRE	8	0.936	0.959	0.974
		GMAT	2	0.911	0.935	0.953
Human/Human ADJACENT		State wide	18	0.928	0.940	0949
		Toefle	4	0.962	0.971	0.977
		GRE	4	0.955	0.971	0.981

K = number of effect sizes; CI = confidence interval; *ES*=effect size

〈Table 6〉 Subgroup analysis for subject characteristics

Category	Sub-group	Sub-outcomes	*k*	-95%CI	ES	+95%CI
AES/Human EXACT	Scorer Expertise	Expert	12	0.570	0.622	0.671
		Human	86	0.489	0.503	0.518
Human/Human EXACT		Human/Expert	12	0.491	0.511	0.532
		Human/Human	27	0.489	0.513	0.536
AES/Human ADJACENT		Expert	14	0.945	0.969	0.983
		Human	53	0.919	0.934	0.946
Human/Human ADJACENT		Expert	14	0.935	0.948	0.959
		Human	20	0.906	0.934	0.955
AES/Human EXACT	School Level	K-12	66	0.487	0.506	0.525
		Undergraduate	12	0.489	0.530	0.571
		Graduate	14	0.456	0.517	0.544
Human/Human EXACT		K-12	30	0.476	0.494	0.512
		Undergraduate	5	0.560	0.581	0.601
		Graduate	4	0.494	0.542	0.590
AES/Human ADJACENT						
		K-12	38	0.919	0.935	0.948
		Undergraduate	10	0.919	0.949	0.969
		Graduate	10	0.934	0.955	0.970
Human/Human ADJACENT						
		K-12	23	0.919	0.933	0.945
		Undergraduate	5	0.956	0.967	0.935
		Graduate	4	0.955	0.971	0.981

K = number of effect sizes; CI = confidence interval; *ES*=effect size

〈Table 7〉 The results of fixed-effects regression analysis by publication year

Category	Standard Parameter	Estimate	Error	z-value	p-value
AES/Human	Intercept	-38.135	4.33	-8.80	0.000
Exact agreement	Publication year	0.019	0.002	8.82	0.000
AES/Human	Intercept	-56.695	8.89	-6.38	0.000
Adjacent agreement	Publication year	0.030	0.004	6.67	0.000

국문요약

에세이 자동 채점 프로그램과 사람 채점자 간 일치도에 관한 메타분석[2)]

신 인 수
(전주대학교 조교수)

에세이 자동 채점은 컴퓨터 기술을 이용한 작문 채점이다. 사람 채점과 자동 채점 간의 일치도에 대한 타당화 연구가 많이 진행중이다. 아직까지 사람들 간의 채점의 일치도에 비해 부족하다는 비판도 있지만, 많은 연구들은 사람들 간의 채점만큼의 타당성을 주장하고 있다. 이 메타분석의 목표는 에세이 답안 채점에서 사람 채점자 간 일치도와 기계 채점과 사람 채점 간의 일치도를 비교해 보는 것이다.

이 메타분석에서 효과크기는 자동 채점 프로그램과 사람 채점자와의 일치율을 랜덤효과모형을 이용하여 추정하였다. 사람들 간 54%의 완전일치도와 비교하여 자동 채점 프로그램과 사람 채점자 간 완전일치도는 52%였다. 사람들 간 94%의 근접일치도와 비교하여 자동 채점 프로그램과 사람 채점자 간 근접일치도는 93%이다. 이 메타분석은 자동 채점 프로그램과 사람 채점자 간 일치도가 사람 채점자들 간의 일치도와 비교될 만큼 매우 높다는 것을 보여주었다. 이 연구는 또한 출판 여부, 자동 채점 프로그램 유형, 에세이 유형, 시험 유형, 전문가 채점 여무, 국가별, 학교 수준별과 같은 연구 특징 변수별 유목 간 일치도 차이를 비교하였다. 끝으로 이 연구를 통해 에세이 자동 채점 프로그램의 개발 및 적용의 시사점과 향후 연구 방향, 이 연구의 한계에 대해 논의 및 제시하였다.

주제어: 에세이 자동 채점, 일치도, 메타분석, 효과크기

2) 이 논문(저서)은 2012년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (NRF-2012S1A5A8023737).