English Teaching, Vol. 78, No. 3, Autumn 2023, pp. 105-124
DOI: https://doi.org/10.15858/engtea.78.3.202309.105
http://journal.kate.or.kr

Meta-Analysis of Inter-Rater Agreement and Discrepancy Between Human and Automated English Essay Scoring*

Jiyeo Yun**

Yun, Jiyeo. (2023). Meta-analysis of inter-rater agreement and discrepancy between human and automated English essay scoring. *English Teaching*, 78(3), 105-124.

Studies on automatic scoring systems in writing assessments have also evaluated the relationship between human and machine scores for the reliability of automated essay scoring systems. This study investigated the magnitudes of indices for inter-rater agreement and discrepancy, especially regarding human and machine scoring, in writing assessment. The mean of the overall population correlation between automated and human scoring in essay writing was .78. The overall common *d* effect size was 0.001. Results from this meta-analysis indicated a strong relationship with no discrepancies between automated and human scoring. Both the I^2 and Q values suggested that the population correlation values studied seemed to be heterogeneous, in contrast to homogenous *d* effect sizes. Therefore, it is necessary to investigate the sources of the between-studies variations for *r* correlations. Practical implications for ways of reporting results of automatic-scoring systems research and limitations of the study are also discussed.

Key words: inter-rater agreement and discrepancy, automated essay scoring, meta-analysis

*This work is extracted and modified from Yun's doctoral dissertation.

**Author: Jiyeo Yun, Elementary School Teacher, Jeongdong Elementary School, Gyeongsangnamdo Office of Education; 25, Daegok 1 gil, Sacheon-si, Gyeongsangdam-do 52523, Korea; yjymodu@korea.kr

Received 30 June 2023; Reviewed 17 July 2023; Accepted 11 September 2023

1. INTRODUCTION

Since the earliest use of computer systems to grade high-school students' essays in the 1960s (Page, 1966), language assessment agencies, with help from a variety of experts in linguistics, education, and computer programming, have developed systems that can score writings. Scoring systems used in language assessment use computer technologies to evaluate and score constructed-response tasks such as essay writings in which examinees express their ideas or responses (Ramineni & Williamson, 2013; Shermis & Burstein, 2003; Williamson, Xi, & Breyer, 2012; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). This process is called automated essay scoring (AES) or computer-automated scoring (CAS). From the beginning phase of the development of AES systems, developers have implemented two stages – a calibration or training stage, and a validation or scoring stage. In the training stage, developers used training sets of essays scored by human raters to build their own scoring models (Attali & Burstein, 2006; Breyer et al., 2014; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Shermis, Koch, Page, Keith, & Harrington, 2002). After the training stage, they have measured agreement and discrepancy between machine scores and 1) scores assigned by humans of the same construct, 2) other machine scores of the similar constructs, and 3) other scores produced by humans or other machine of the similar construct.

Initial studies on automated essay scoring (Burstein et al., 1998; Foltz, Laham, & Landauer, 1999; Landauer, Laham, Rehder, & Schreiner, 1997; Page & Petersen, 1995) focused on comparing correlations between human and machine scores with correlations between two human raters' scores, because agreement between human and machine scores has been a means to detect the quality of automated essay scoring (Williamson et al., 2012). Specifically, some initial studies on AES systems have focused on either reliability or validity issues. Foltz et al. (1999) performed research to evaluate the reliability of AES, by scoring essays from GMAT (Graduate Management Administration Test) analytical writing assessments, and reported high correlations between scores produced by humans and machines. Meanwhile, validity of AES has been evidence of comparing of scores produced by AES systems with those of humans from the initial implementation. Attali and Burstein (2006) compared scores assigned by AES, individual humans, and average scores from two human raters. They investigated multiple essay scores by received the same examinees in K-12 writing assessments to show validating AES. As research progressed, researchers moved their concentration to comparing machine scores with other criteria that measure similar constructs, which could be seen as efforts to infer validity. Validation of AES might include several different aspects such as assessment attributes, language features, algorithms, and so on (Ifenthaler, 2022).

Therefore, a focus of the study is to evaluate inter-rater agreement and discrepancy, especially reliability of writing assessments in which two or more raters have participated.

With regard to reliability coefficients, inter-rater agreement includes measures of absolute accuracy such as consensus or agreement in scores, and measures of directional similarity in two sets of scores produced two raters (Brown, Glasswell, & Harland, 2004; Cohen, Ben-Simon, & Hovav, 2003; LeBreton & Senter, 2008; Ramineni, 2013; Weigle, 2011). Research has shown various results regarding agreements including similarity and discrepancies between human and machine scores. With respect to agreements, researchers have used percentages, correlations, and kappa to report the agreement between human and machine scores (Attali, 2007; Breyer et al., 2014; Deane, Williams, Weng, & Trapani, 2013). On the other hand, the standardized mean difference has been used to display discrepancies between human and machine scores (Jin & Park, 2012; Ramineni, 2013).

Scoring systems using computers may not totally replace human raters, but at least have played a role of a second rater – checking writing quality, or helping to reduce time and grading resources. Individual studies have reported a diverse set of results to show evidence that automatic scoring systems are reliable and valid. They have provided agreement coefficients including correlation, as well as standardized mean differences as indices of discrepancies in order to support effectiveness of machine scoring (Streeter, Psotka, Laham, & MacCuish, 2002). However, a variety of results regarding the relationships among indices cause obscure relationships to understand. Meta-analyses enable the results including wide ranges and unclear relationships across studies to be comparable and then them to be summarized. In addition, meta-analyses might make it possible to focus to the magnitude of effects across studies. This raises the question of what relationships exist between measures of agreement and discrepancy.

The main purpose of the study is to investigate the magnitudes of indices for inter-rater agreement and discrepancy specifically in regard to human and machine scoring in writing assessment. The task of the study is to figure out the following three research questions using existing studies.

1) To what extent is agreement and discrepancy between human and machine scoring in writing assessments?
2) Are the degrees of agreement and the discrepancies consistent across studies?
3) Do numbers of point on the scales influence the rates of agreement and discrepancy?

2. REVIEW OF LITERATURE

This chapter begins with the features of automated scoring systems, and includes features of inter-rater agreement and discrepancy indices. Finally, it briefly describes meta-analysis

models used to integrate the results from studies of inter-rater agreement and discrepancy and to estimate the magnitudes of effect sizes from empirical studies.

2.1. Automated Scoring Systems

Ramineni and Williamson (2013), as well as Shermis (2014) stated that four major AES systems have been widely used and commercialized: *e-rater* (by Educational Testing Services), Intelligent Essay Assessor (IEA, by Pearson Knowledge Technologies), IntelliMetric (by Vantage Learning), and Project Essay Grade (PEG, by Measurement Incorporated). Current AES systems are trying to detect and examine essay features related to essay content (e.g., vocabulary, sentence structure, organization, and style), rather than scoring correctness of the essay including grammatical errors, spelling, punctuation, and so on (Valenti, Neri, & Cucchiarelli, 2003). However, each AES system focuses on either surface features related to English usage, lists of words, essay lengths, and so on, or deep features involving content, structure, and organization, etc. (Dikli, 2006; Hearst, 2000; Yang et al., 2002). For instance, IEA evaluates mainly meaning and content using Latent Semantic Analysis (LSA), whereas *e-rater* and IntelliMetric analyze semantic, syntactic, and discourse levels of essays employing Natural-Language Processing (NLP) techniques (Koul, Clariana, & Salehi, 2005; Warschauer & Ware, 2006). In other words, each scoring system uses different methods that are based on LSA, NLP, or statistical methods. Specifically, IEA scores essays using LSA that uses matrix-algebra techniques, whereas *e-rater* and IntelliMetric have employed an NLP technique that is an Artificial Intelligence (AI) application to extract measurable features (i.e., frequencies of words, weights of words, etc.) corresponding to attentive constructs such as syntactic and discourse structures, lexical complexity and so on.

Developers have devised and updated the systems' general models, and variant models such as prompt-specific models and weighting-feature models (Attali, Bridgeman, & Trapani, 2010; Ben-Simon & Bennett, 2007; Ramineni, 2013). Simply the basic concept of building models involves extracting measurable writing features or semantic similarities from numerous essays pre-scored by humans. At the same time, as Attali and Burstein (2006) stated, machine scores are produced by using an appropriate scaling such as setting the same mean and standard deviation for machine scores as those of scores assigned by human raters from large samples of essays. Due to the process of building models in which AES systems have been trained by numerous pre-scored sample essays through scaling and predicting process, AES systems might emulate human scoring. Like human raters, automated scoring models might learn and show these kinds of rater variability, because the models are based on human scoring patterns.

2.2. Inter-Rater Agreement and discrepancy

Test-score reliability refers to the ratio of the true-score variance to the observed-score variance in classical test theory (CTT), but it is possible to estimate reliability in various ways. In addition to three popular ways (i.e., test-retest reliability, alternative-form reliability, and internal consistency reliability), inter-rater reliability is commonly used to estimate the reliability of essay scoring (Burry-Stock, Shaw, Laurie, & Chissom, 1996). For estimating reliability in automatic essay assessments, some researchers have investigated internal consistency reliability, providing alpha coefficients and Spearman-Brown estimates of each set of scores – human and machine (Chodorow & Burstein, 2004; Sireci & Rizavi, 2000). Others have evaluated inter-rater reliability of human and machine scores (Burstein et al., 1998; Shermis, 2014). Most investigators of studies on automatic scoring systems have reported percentages of agreement, kappa, correlation, and standardized mean differences (using means and standard deviations) to show both agreement and discrepancy as evaluation criteria for the association of machine scores with scores assigned by human raters (Breyer et al., 2014; Deane et al., 2013; Shermis, 2014).

2.2.1. Correlation

As mentioned earlier, initial studies on AES systems investigated human-machine score correlations and human-human score correlations, so correlation coefficients have been a common outcome measure for the research on AES. When reporting correlations, most researchers used Pearson's correlation, but a few investigators reported Spearman's correlation (Landauer et al., 1997; Li, Link, Ma, Yang, & Hegelheimer, 2014; Sireci & Rizavi, 2000; Wang & Brown, 2007). The main reason for using Spearman's rho correlation in AES research might be due to the fact that rating rubrics have rank-order features. Therefore, one uses Pearson's correlation (r) to analyze the degree of linear association between two variables of interval or ratio data, while one uses Spearman's rho correlation (r_s) to examine the magnitude of association between two variables of ordinal data. Regardless of levels of measurement (e.g., ordinal and interval), data with large samples produce similar results in terms of Pearson and Spearman correlations (Bishara & Hittner, 2012). Pearson's sample correlation coefficient between human and machine scores can be calculated as the covariance of human and machine scores is divided by the product of the standard deviations of human and machine scores.

2.2.2. Standardized mean difference

Studies reporting d as an effect size have addressed the extent to which machine scores

differ from human scores on average. Researchers have argued for a standardized mean difference of 0.15 or less as a cut-point value or threshold, because researchers consider *d* effect sizes of 0.15 or smaller to be trivial (Ben-Simon & Bennett, 2007; Breyer et al., 2014; Ramineni, 2013; Rudner, Garcia, & Welch, 2006). In addition, Williamson et al. (2012) suggested the standardized mean difference between two human raters of 0.15 or less as a standard. Therefore, a larger *d* effect size in AES research indicates that human scoring differs from machine scoring on average. Smaller *d*s or *d*s close to zero suggest no differences between two scorings on the average.

Most researchers have presented descriptive statistics on human and machine scoring using means and standard deviations as an initial analysis. Based on the statistics, one can calculate standardized mean differences to evaluate the degree of discrepancies and tendency for the scores generated by machine scoring to be higher (or lower) than the scores assigned by human raters. In addition to the size of the *d* effect-size index, the direction of the effect size helps to identify whether one scoring system is stricter than another is. Suppose one computes *d* by subtracting the machine-score mean from the human-score mean. If the direction is positive, human scoring is more lenient than machine scoring. Meanwhile, a negative sign indicates that machines give higher scores to essays than human raters do.

2.3. Meta-Analysis Models

Effect size refers to magnitude of some effect; effect sizes may measure mean differences, relationships between variables, and proportion differences (Borenstein, Hedges, Higgins, & Rothstein, 2009). Meta-analyses examine whether studies within a collection share a common true effect size or display an average effect size from a distribution of effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2010). Generally, meta-analyses use techniques that assign different types of weights such as fixed-effect weights or random-effect weights to the selected studies, which enables estimating the weighted mean. Based on the evaluation criteria of Williamson et al. for automated scoring systems (2012), the effect-size indices indicating the relationship between human and machine scores (i.e., agreement or discrepancy) can be classified into two indices (i.e., correlation and standardized mean differences) estimating inter-rater agreement.

In order to analyze the effect sizes, one uses two types of models, fixed-effect models and random-effect models. They are distinguished depending on what weights are allotted. Under the fixed-effect models, one estimates the common true effect size assuming only sampling error within studies exists. Meanwhile, under random-effect models, one can estimate the average effect size assuming each study has its own true effect size, and sampling errors within studies as well as between studies variation exist. To be specific, simple sampling error, reflected in the within-study variances (i.e., σ_i^2) depends on the

primary sample sizes of the study, while another kind of sampling error, between-studies variances (i.e., τ^2 or σ_θ^2), arises from variations across studies (Schmidt, Oh, & Hayes, 2009). Therefore, sample sizes are less likely to affect the estimates in random-effect models.

Before selecting fixed- or random-effect models, researchers may examine whether all effect sizes within a meta-analysis come from a common population through the heterogeneity test (Schmidt et al., 2009). The null hypothesis for the *Q* test of heterogeneity is that the between-studies variance equals to zero, which is denoted as

$$H_0: \theta_1 = \theta_2 = \ldots . \theta_k,$$

where θ_i is the population effect size in the study *i* for *i* = 1 to *k* studies. For this study, θ represents $Z(\rho)$ and δ, which indicate the true effect size of correlation, and standardized mean difference, respectively. The *Q* statistic is sensitive to the number of studies, so Higgins, Thompson, Deeks, and Altman (2003) suggested an index indicating the amount of heterogeneity (i.e., I^2). Moreover, Higgins et al. (2003) pointed out that I^2 is comparable across meta-analysis studies that use different study sizes and different outcomes.

3. METHODOLOGY

The procedure of meta-analyses consists of data gathering with search procedures and criteria for inclusion and exclusion, data cleansing, and data analysis.

3.1. Data Gathering

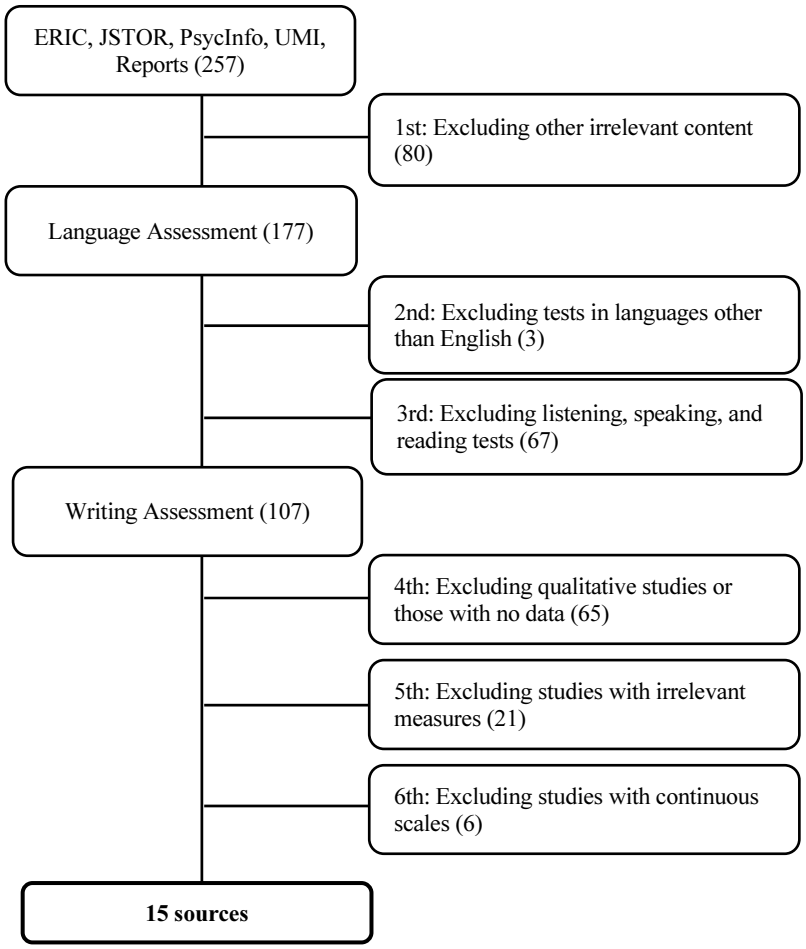
To search for studies on the inter-rater agreement of human and machine scoring in writing assessments, terms including ‘automated scoring’, ‘automatic scorer’, ‘automatic scoring’, ‘computer rater’, ‘human rater’, ‘human scoring’, and ‘language assessment’ were used. Lists showing titles from digital libraries and online databases such as JSTOR, ERIC, PsycINFO, ProQuest, and Google Scholar were scanned. Moreover, relevant journals including *Applied Measurement in Education, Journal of Educational Measurement, Journal of Information Technology Education*, *Language Testing*, and *Assessing Writing*, and the archives of Educational Testing Services and Vantage Learning reports were scanned. After titles, abstracts, and references of each study were reviewed, 257 papers in the initial check were obtained.

Several inclusion and exclusion criteria were used to select studies that report appropriate outcomes for inter-rater agreement between automated and human scoring. Figure 1 shows the search procedures. First, studies must have tested language as the subject field, and those from other areas (e.g., math, science, medical, etc.) were excluded; eighty studies out of 257 initial findings did not meet the criterion thus were excluded. Second, three studies that

assessed French or Spanish languages were excluded. Third, sixty-seven more studies were excluded because the studies assessed writing skills rather than listening, speaking, and reading skills. Next, sixty-five out of 107 writing-assessment studies that were commentaries and instructional-practice articles or qualitative studies were excluded, because they did not appropriate to conduct the meta-analysis. Specifically, studies offering information on correlations and standardized mean differences were selected. Therefore, twenty-one studies that did not have sufficient outcome measures were excluded. Furthermore, six studies using continuous score scales rather than categorical scales were excluded. In sum, fifteen study sources with seventy-one samples of outcomes had information on both correlations and standardized mean differences (i.e., two sets of seventy-one effects).

FIGURE 1

Search Procedure



3.2. Data Cleansing

In order to do data cleaning and proper coding for the meta-analysis, two coders individually coded the studies for two aspects: test and effect-size characteristics. Regarding test features, the coders coded application context, AES system used, and name of test used in the primary research. For the effect-size characteristics, two coders coded the number of points on the scales, sample sizes, the number of studies, and the types of effect sizes. Percentages of agreement among coders to determine inclusion or exclusion judgments for fifteen papers were calculated. Coders agreed on from 79.8% to 100% of judgments, and then discussed all disagreements until they were fully resolved. A dataset with two indices (i.e., correlations and *d*-effect sizes) were made. These effect sizes and their variances were calculated or transformed; the correlations were transformed to Fisher-*z* statistics.

3.3. Data Analysis

For the overall meta-analysis models, appropriate analyses were conducted using `rma`, `robust`, and `robu` functions in the *metafor* (Viechtbauer, 2010) and *robumeta* packages (Fisher & Tipton, 2015) in *R* software Version 3.3.2 (R Core Team, 2016). Before selecting the most appropriate models (i.e., fixed- or random-effect models), tests of heterogeneity were conducted. Moreover, the existing studies for this meta-analysis include multiple outcomes, which seems to be dependent. Assuming no correlation or perfect correlation between effect sizes within a study can lead to either under- or over-estimates of the variance of the effect sizes (Scammacca, Roberts, & Stuebing, 2014). Therefore, in order to adjust the standard errors of all measures, robust variance estimation (RVE) techniques were used. Besides, for further investigations of whether between-studies differences in each effect size is associated with differences in numbers of point on the scales, moderator analyses using numbers of point on the scales were conducted.

4. RESULTS

Magnitudes of overall effects were evaluated for the two indices (i.e., correlation and standardized mean difference) in the dataset. In addition, the group mean effect sizes for scales with different numbers of scale points were compared.

4.1. Description of Research Sources

Table 1 summarizes fifteen research sources selected to build the dataset. It contains

information about application contexts, the types of AESs and the tests that were scored, numbers of scale points, sample sizes, the number of effect sizes per source, types of effect sizes reported, and year published.

TABLE 1

Description of the Fifteen Studies

ID	Author	Year	Con-text	AES[a]	Test Name	# of Scale points	Sample sizes	# of studies	ES[b] Types
1	Attali	2007	H	*e*-	TOEFL	5	5,006	4	*r, d*
2	Attali et al.	2010	H	*e*-	TOEFL, GRE	5, 6	26 to 139	3	*r, d*
3	Breyer et al.	2014	H	*e*-	GRE	6	748, 773	2	*r, d, p*
4	Bridgeman et al.	2012	H	*e*-	TOEFL, GRE	5, 6	2,925 to 103,465	14	*r, d*
5	Burstein & Chodorow	1999	H	*e*-	TOEFL	6	562, 576	2	*r, d, p*
6	Deane et al.	2013	H	*e*-	CBAL[c]	5	448 to 1,539	5	*r, d, K, p*
7	Doğan	2014	L	*e*-	Essay	6	49	1	*r, d*
8	Jin & Park	2012	L	In-	Essay	4, 5	218	4	*r, d, p*
9	Koul et al.	2005	L	In-	Essay	5	66	1	*r, d*
10	Powers et al.	2002	H	*e*-	GRE	6	63	1	*r, d, K, p*
11	Ramineni	2013	L	*e*-	Essay	6	411 to 450	4	*r, d*
12	Rudner et al.	2006	H	In-	GMAT	6	500	6	*r, d, p*
13	Shermis	2014	L	*e*-, In-, I, P	Essay	3, 4, 6	568 to 601	20	*r, d, K, p*
14	Sireci & Rizavi	2000	H	*e*-	WritePlacer Plus	4	464, 467	2	*r, d, K, p*
15	Weigle	2011	H	*e*-	TOEFL	6	376	2	*r, d, p*

Note. Context: H = High-stakes test, and L = Low-stakes test

[a]AES = Automated Essay Scoring system. The *e*-, In-, I, and P dente *e*-rater, IntelliMetric, IEA, and PEG, respectively.

[b]ES = Effect Size. The *r, d, K,* and *p* are abbreviations of correlations, standardized mean differences, kappa, and percentage agreements, respectively.

[c]The Cognitively Based Assessment of, for, and as Learning (CBAL) English Language Arts assessment.

To be specific, these sources were reports, articles published in journals, or papers presented at conferences between 1999 and 2014. AES systems have usually applied in one of two contexts; high-stakes and low-stakes test contexts. Some researchers examined essays

from GMAT analytical writing assessment, GRE analytical writing assessment, TOEFL writing, the WritePlacer Plus test, and the Cognitively Based Assessment of, for, and as Learning (CBAL) English Language Arts assessment, which were considered to be high-stakes tests. The results of such high-stakes tests often play a role in important decisions such as whether to admit candidates to academic programs.

On the other hand, some investigators examined essay prompts used to give feedback to students in the classroom, or used as a means of measuring students' achievement and identifying their learning processes. These were deemed to be low-stakes tests. Furthermore, in all sources the scores assigned by human raters and one or more types of AES system (i.e., *e-rater*, IntelliMetric, IEA, and PEG) were used to investigate automated-scoring-system effectiveness. The categorical score scales had used three-, four-, five-, and six-point scales. Among fifteen research reports, three sources reported only one study and the others reported more than one study. Moreover, fifteen sources reported both *r* and *d* ES indices.

Among the seventy-one effect sizes in the dataset, two sources reported both Pearson and Spearman correlations, but the other sixty-nine sources reported only Pearson correlations. Moreover, thirty-four sources reported *d* effect sizes directly. The others reported means and standard deviations for human and machine scores, which allowed researchers to compute *d* effect sizes. Except for four sources, the other sources used sample sizes of over one hundred. Furthermore, the sample sizes in the dataset ranged from 26 to 103,465. Correlations ranged from .42 to .95 with an unweighted mean of .766, and standardized mean differences had an unweighted mean of 0.021 and ranged from -0.19 to 0.28.

4.2. Heterogeneity Tests

First of all, heterogeneity tests were conducted. Table 2 displays the results of heterogeneity tests for each effect. The Q values and I^2 statistic for the Fisher-*z* transformation were highly significant (Q (70) = 9590, I^2 = 99.5%, $p < .001$), whereas the Q value for *d* effect sizes was not significant (Q (70) = 1.9, I^2 = 0%, p = 1.0). A significant Q test of heterogeneity suggests that the effect sizes within a meta-analysis are not from one population, which means that a random-effect model is appropriate.

TABLE 2

Heterogeneity Tests of Correlation and *d*

Effect-size index	Sample size (*n*)	# of effect sizes	Indices of heterogeneity		
			Q	$\hat{\tau}^2$ ($\hat{\tau}$)	I^2 *(%)*
Fisher-*z* of correlation	473278	71	9590***	0.035 (0.187)	99.5
d	*473278*	*71*	*1.9*	*0*	*0*

Note. ***$p < .001$

To be specific, both the I^2 and Q values suggested that the variance of the population ρ values did not equal zero, which meant that the population studied seemed to be heterogeneous. On the other hand, the variance of the population δs equaled zero, which indicated homogeneous *d* effect sizes. Consequently, it was appropriate to conduct random-effects analyses for *r*, in contrast to the fixed-effect model that was applicable to the *d* effect sizes.

4.3. Overall Means of the Correlation and *d*

The first research question is to investigate the extent of agreement and discrepancy between human and machine scoring in writing assessments. The mean of the hierarchical-effects model for the overall *r* between automated and human scoring in essay writing was .78 with a 95% confidence interval ranging from .74 to .81, seen in Table 3. On the other hand, according to the results of the fixed-effect model for the *d* effect sizes, the overall common effect was 0.001 with a 95% confidence interval between -0.0152 and 0.0173. Therefore, the results show .78 of agreement rates and 0.001 of discrepancy rates as answers of the first research question. According to Cohen's conventions (Cohen, 1992), the correlations were all quite large, which means that machine scoring is strongly related with human scoring.

TABLE 3

Effect Sizes Indices from Hierarchical Weighted Model

Effect-size index	Estimate[a]	Robust SE	$\hat{\tau}^2$	$\hat{\omega}^2$	95% CI[b]	*k*[c]
Fisher-*z* (*r*)	1.04** (.78)	0.037	0.014	0.02	0.95, 1.13 (.74, .81)	15 (71)
d [d]	0.001	0.0083	0	2.103	-0.0152, 0.0173	15 (71)

Note. [a]Back transformed estimates are in parentheses.
[b]Back transformed 95% CI is in parentheses.
[c]*k* indicates numbers of studies, and numbers of effect are sizes in parentheses.
[d]The estimate of the mean *d* effect size, its standard error, and its confidence interval came from the Fixed-effects model.
**$p < .01$

4.4. Consistency Across Studies

The second research question is to examine whether the degree of agreement and discrepancy is consistent across studies. Both the larger Q statistics (Q (70) = 9590) and the larger I^2 (99.5%) indicate more heterogeneity seen in Table 2. Additionally, the tau-squared estimate for the Fisher-*z* transformation values is 0.035, which means that the *r* correlations had some between-studies variations. On the other hand, given that the structure of effect

sizes collected was hierarchical, the *d* effects had some variation between effects within studies; the estimate of their within-study variation is $\widehat{\omega}^2$ = 2.1, seen in Table 3. This suggested that the *d* effects had relatively large between-effects variation within studies compared to the between-studies variation. In other words, the results show the degree of discrepancies is consistent across studies, but the degree of agreement is not.

4.5. Influence of the Numbers of Point on the Scale

The third research question is about the numbers of point on the scales. Table 4 displays the results of moderator analyses using the numbers of points on the scales (i.e., 3-, 4-, 5-, and 6-point scales) for correlations and standardized mean differences. However, no significant differences were found by number of scale points for correlations and standardized mean differences seen in Table 4. Therefore, numbers of point on the scales do not influence the rates of agreement and discrepancy specifically correlations and *d* effects.

TABLE 4

Moderator Analyses for Correlation and *d*

Index	Q_{between} (df = 3)	Scale points (k)	$\overline{\text{ES}}$[a]	*SE*
Correlation	1.381	3-point (8)	.757***	.069
		4-point (12)	.774***	.056
		5-point (23)	.772***	.041
		6-point (28)	.790***	.037
D	0.12	3-point (8)	.0058	.167
		4-point (12)	-.0008	.023
		5-point (23)	.0002	.010
		6-point (28)	.0098	.366

Note. *k* is the number of effect sizes.
SE = standard error
[a]$\overline{\text{ES}}$ denotes effect-size estimate.
***p < .001

5. DISSCUSSION AND CONCLUSION

Automated essay scoring systems have a large advantage in reducing time and cost, especially when the number of examinees is excessive scale, to score construct response items. The purpose of the study was to investigate the magnitudes of inter-rater agreement and discrepancy between human and automated essay scoring in writing assessments. This chapter begins with a summary of the results for the study in terms of answering the research questions including practical implications for ways of reporting results of automatic-scoring systems research. Additionally, further works and limitations of the study are also discussed.

First of all, the first research question in the study was to evaluate the extent of agreement and discrepancy between human and machine scoring. Based on the results from existing studies, the overall mean effects were .78 for correlations, and 0.001 for *d* effect sizes. Compared to the evaluation criteria suggested by Williamson et al. (2012), Ramineni and Williamson (2013), Deane (2013), as well as Rotou and Rupp (2020), the degrees of inter-rater agreement seen in this study are above current thresholds for correlations. To be specific, the mean correlation of .78 is larger than the suggested correlation of .70 for minimum adequate agreement. Moreover, the standardized mean difference of 0.001 from the study are much smaller than the 0.10 or 0.15 cutoff suggested for standardized mean differences between human and machine scores. The results of agreement and discrepancy between human and machine scoring show levels of agreement and discrepancy that met the criteria. Furthermore, since Yun (2017) suggested correlations of .75 as the standard cutoff, this study proposes that the correlation coefficient should be above .75 to confirm inter-rater reliability in the AES research area.

Some researchers have written systematic literature reviews on AES systems introducing features of each AES such as techniques used in AES, models built, and so on. They have also described the strengths and weaknesses of each AES, as well as the methods and results of empirical studies they collected (Blood, 2011; Hussein, Hassan, & Nassef, 2019; Ifenthaler, 2022; Klebanov & Madnani, 2020; Ramesh & Sanampudi, 2022). However, a major strength of this study is that it provides acceptable minimum correlations and *d* effect sizes between human and automated scoring, resulting in a unique meta-analysis.

The second research question in the study was to investigate whether the degree of agreement and discrepancy is consistent across studies. Based on the results, the *d* effects' within-study variation (i.e., $\widehat{\omega}^2 = 2.1$) indicates that the *d* effects had relatively large within-study differences (i.e., study-specific variations) compared to the between-studies differences. On the other hand, the tau-squared estimate for the Fisher-*z* transformation value is 0.035, suggesting the *r* correlations had some between-studies variations. In short, the degree of discrepancy is consistent across studies while the degree of agreement is not. Researchers need to further investigate sources of the between-studies differences for *r* correlations.

Moderator analyses using point values on the scales (i.e., 3-, 4-, 5-, and 6-point scales) were conducted to explain the between-studies variances for each index to answer the third research question. However, correlation and standardized mean differences did not have any significant differences across points on the scales. According to a study by Jeong and Lee (2016), correlations among the results using different points were extremely high, which indicates negligible differences among the results employing different scales. Rotou and Rupp (2020) also show the correlations do not depend on the scales. The results of moderator analyses of the present study are consistent with Jeong and Lee (2016), and Rotou and Rupp

(2020).

However, since the priority of the present study is to investigate the magnitudes of agreement and discrepancies between human and automated scoring in writing assessments as inter-rater reliability, this study did not strongly consider linguistic features, including different dimensions of writing quality, as well as issues generated in writing assessments such as validity, rater variability, group fairness, etc. Therefore, this paper suggests further works related to these limitations. First, Ifenthaler (2022) stated that correlations between human and automated scoring is sufficiently high, but agreement rates including exact and adjacent agreement do not come close to proper rates. Thus, other agreement indices in examining inter-rater reliability might show different levels of agreement. Further research regarding AES systems should use additional measures of agreement; such as percentages of exact and adjacent agreement, and kappa statistics including quadratic-weighted kappa (QWK) in order to investigate inter-rater reliability (Ke & Ng, 2019; Rotou & Rupp, 2020).

Second, as Ifenthaler (2022) pointed out, studies reporting correlations between human and automated essay scoring in AES systems have mainly applied high-stakes test contexts instead of in low-stakes test contexts. Moreover, Uzun (2018) also reported a lower correlation (r_s = .246) between human and automated essay scoring in classroom assessments due to small sample sizes. Further investigations with moderator analyses, such as assessment contexts, might trace sources of between-studies variations for correlations.

Lastly, since primary studies in this work were published from 1999 to 2014, it seems somewhat outdated given rapid technological change. As Ramesh and Sanampudi (2022) pointed out, as AES systems began employing new technologies and approaches, such as natural language processing, future works should include studies on recent AES systems, and focus on different writing features that were not included and analyzed in this study.

Applicable levels: Secondary, tertiary

REFERENCES

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

*Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS RR-07-21). Princeton, NJ: ETS.

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

*Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment, 10*(3), 1-15.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning, and Assessment, 6*(1), 1-17.

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17*(3), 399-417.

Blood, I. (2011). Automated essay scoring: A literature review. *Studies in Applied Linguistics and TESOL, 11*(2), 40-64.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, England: John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effectsmodels for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.

*Breyer, F. J., Attali, Y., Williamson, D. M., Ridolfi-McCulla, L., Ramineni, C., Duchnowski, M., & Harris, A. (2014). *A study of the use of the e-rater scoring engine for the analytical writing measure of the GRE revised general test* (ETS RR-14-24). Princeton, NJ: ETS.

*Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27-40.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004) Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105-121.

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater agreement indexes for performance assessment. *Educational and Psychological Measurement, 56*(2), 251-262.

*Burstein, J. & Chodorow, M. (1999, June). Automated Essay Scoring for nonnative English speakers. *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing* (pp. 68-75). College Park, MD: Association for Computational Linguistics.

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

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater®'s performance on TOEFL® essays* (TOEFL Report 73, RR-04-04). Princeton, NJ: ETS.

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

Cohen, Y., Ben-Simon, A., & Hovav, M. (2003). *The effect of specific language features on the complexity of systems for automated essay scoring.* Paper presented at the 29th Annual Conference of the International Association for Educational Assessment, Manchester, England.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24.

*Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *The Journal of Writing Assessment, 6*(1), 1-15.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-36.

*Doğan, A. (2014). Automated essay scoring versus human scoring: A reliability check. In A. Akbarov (Ed.), *Linguistics, culture and identity in foreign language education* (pp. 277-288). Sarajevo, Bosnia and Herzegovina: IBU Publications.

Fisher, Z., & Tipton, E. (2015). *Robumeta: Robust Variance Meta-Regression (version 1.6)*. Retrieved on January 17. 2017, from http://cran.r-project.org/web/packages/robumeta.pdf

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939-944.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22-37.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560.

Hussein, M., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208.

Ifenthaler, D. (2022). Automated essay grading systems. In O. Zawacki-Richter & I. Jung (Eds.), *Hanboock of open, distance and digital education* (pp. 1-15). Singapore: Springer Nature Singapore.

Jeong, H., & Lee, W. (2016). The level of collapse we are allowed: Comparison of different response scales in Safety Attitudes Questionnaire. *Biometrics & Biostatistics International Journal, 4*(4), 128-134.

*Jin, K., & Park, T. (2012). Exploring automated scoring of Korean high school students' English composition. *Multimedia-Assisted Language Learning, 15*(1), 11-29.

Ke, Z., & Ng, V. (2019, August). Automated essay scoring: A survey of the state of the art. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300-6308). Macao, China.

Klebanov, B. B., & Madnani, N. (2020, July). Automated evaluation of writing–50 years and counting. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7796-7810). Stroudsburg, PA: Association for Computational Linguistics.

*Koul, R., Clariana, R. B., & Salehi, R. (2005). Comparing several human and computer-based methods for scoring concept maps and essays. *Journal of Educational Computing Research, 32*(3), 227-239.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without word order?: A comparison of latent semantic analysis and humans. In M. G. Shafto, & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66-78.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan, 47*(5), 238-243.

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

*Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*, 103-134.

R Core Team. (2016). *R: A language and environment for statistical computing* [computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review, 55*, 2495-2527.

*Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing, 18*, 40-61.

Ramineni, C., & Williamson, D. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25-39.

Rotou, O., & Rupp, A. A. (2020). *Evaluations of automated scoring systems in practice* (Research Report No. RR-20-10). Princeton, NJ: ETS.

*Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4). 1-22.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research, 84*(3), 328-364.

Schmidt, F. L., Oh, I. & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62*, 97-128.

*Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76.

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

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

*Sireci, S. G., & Rizavi, S. (2000). *Comparing computerized and human scoring of students' essays*. Laboratory of Psychometric and Evaluative Research Report (No. 354). Amherst, MA: School of Education, University of Massachusetts.

Streeter, L., Psotka, J., Laham, D., & MacCuish, D. (2002). The credible grading machine: Automated essay scoring in the DOD. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*. Orlando, FL: I/ITSEC.

Uzun, K. (2018). Home-grown automated essay scoring in the literature classroom: A solution for managing the crowd?. *Contemporary Educational Technology, 9*(4), 423-436.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2,* 319-330.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology, Learning, and Assessment, 6*(2), 1-29.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157-180.

*Weigle, S. C. (2011). *Validation of Automated Scores of TOEFL iBT Tasks against Nontest Indicators of Writing Ability* (TOEFL iBT Research Report -15, RR-1-24). Princeton, NJ: ETS.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

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.

Yun, J. (2017). *The impact of rater variability on relationships among different effect-size indices for inter-rater agreement between human and automated essay scoring*. Unpublished doctoral dissertation, Florida State University, Tallahassee.