Title A systematic review of automated writing evaluation systems
Author(s) Shi Huawei and Vahid Aryadoust

This is a post-peer-review, pre-copy/edit version of an article published in *Education and Information Technologies*. The final authenticated version is available online at: https://doi.org/10.1007/s10639-022-11200-7

Huawei, S., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. Education and Information Technologies. Advance online publication. https://doi.org/10.1007/s10639-022-11200-7

A Systematic Review of Automated Writing Evaluation System

Abstract

This study conducted a systematic review of empirical research on Automated Writing Evaluation (AWE) from 1996 till 2020. Using Scopus, we identified 105 published papers on AWE scoring and coded them within an argument-based validation framework. The major findings are: (i) AWE scoring research had a rising trend, but was heterogeneous in terms of the language environments, ecological settings, and educational level; (ii) a disproportionate number of studies were carried out on each validity inference, with the evaluation inference receiving the most research attention, and the domain description inference being the neglected one, and (iii) most studies adopted quantitative methods and yielded positive results that backed each inference, while some studies also presented counterevidence. Lack of research on the domain description inference combined with the heterogenous contexts indicated that construct representation in the AWE scoring field needs extensive investigation. Implications and directions for future research are also discussed.

Keywords: Automated writing evaluation; argument-based validation; automated essay scoring

A Systematic Review of Automated Writing Evaluation Scoring within the Argument-Based Validation Framework

In writing assessment, automated writing evaluation (AWE) systems, also referred to as Automated Essay Evaluation, Automated Essay Scoring (Hockly, 2019; Warschauer & Ware, 2006) are developed based on a number of research and technological advances such as natural language processing and latent semantic analysis that enable “the process of evaluating and scoring written prose via computer programs” (Shermis et al., 2013, p. 1; also see Shermis & Burstein, 2003). AWE systems were originally developed to generate summative scores for written essays in high-stakes tests. Nevertheless, they were gradually revised and readapted to contribute to classroom instruction and assessment by providing automated feedback as well as scores (Shermis et al., 2016; Stevenson, 2016; Stevenson & Phakiti, 2014).

Conventionally, a distinction has been made between automated scoring and automated feedback (Burstein et al., 2020; Shermis et al., 2016; Stevenson & Phakiti, 2019; Ware, 2011; Xi, 2010). According to Ware (2011, p. 769), automated scoring refers to “the provision of automated scores derived from mathematical models built on organizational, syntactic, and mechanical aspects of writing” for the assessment purposes. Automated feedback, on the other hand, serves the purpose of providing assistance and has a formative goal rather than being a tool exclusively for summative assessment. Although nowadays virtually all AWE systems that support automated feedback also have the feature of providing automated holistic or/and analytical scores (Burstein et al., 2020), the distinction between automated scoring and automated feedback is necessary and justified because of the different purposes the two assessment types serve (Weigle, 2013a) as well as the different validation methods and evidence they require (Chapelle et al., 2015; Xi, 2010).

Research has shown that AWE systems can provide reliable and valid measures of writing abilities (e.g., Keith, 2003; Shermis, 2014). In fact, AWE systems have been used as a second rater to score the writing components of several high-stakes tests such as the Graduate Record Exam (GRE), Test of English as a Foreign Language (TOEFL), and placement tests (Williamson et al., 2012). However, the application of AWE scores in large-scale assessments and in classroom contexts is not without criticisms. One of the first volumes that opposed AWEs is a book edited by Ericsson and Haswell (2006), in which a series of papers questioned the accuracy of AWE systems and objected to their application in assessments.

Overall, most opposing voices have centered on the susceptibility of the systems to “gaming” (cheating the AWE systems into assigning higher scores than human raters) along with the question of scoring and feedback accuracy (e.g., Powers et al., 2002a), and mis-/underrepresentation of the writing construct (e.g., Perelman, 2014; Vojak et al., 2011). Even some agencies expressed their concerns over AWE scoring. For example, in a position statement, the *National Council of Teachers of English* (NCTE) (2013) pointed out that AWE systems missed the social nature of writing and were unable to handle complex writing features. At the very heart of these controversies and criticism, as suggested by Stevenson and Phakiti (2019), are the “paradigmatic differences regarding the nature of writing and writing instruction” (p. 129).

While it might be true that traditional AWE systems typically rely on surface-level language features such as the total number of words per essay, sentence length, etc. and that the construct of writing might not be fully represented in some of these systems, the latest development of AWE systems have managed to incorporate more complex or deep language features (e.g., world knowledge) (Deane, 2013; Shin & Gierl, 2020). These developments indicate the increasing mutual understanding and acceptance between opposing sides.

The Rationale of the Present Study

Although there is an array of studies on AWE scoring, the results have often been mixed (e.g., Cohen et al., 2018). There is a need to evaluate and summarize the findings of previous AWE studies to identify controversies surrounding AWE systems, determine their overall precision and accuracy in different contexts, and make the available evidence accessible. We carried out a review of the relevant research with a focus on validity to synthesize the results of research based on a systematic search, aiming to generate a more comprehensive and objective picture of the field (Li & Wang, 2018).

As previously discussed, automated scoring and automated feedback are distinct. In this review, we chose to focus on the research that investigated AWE scoring due to the differences in nature and usage of the two research streams, as previously discussed (Stevenson & Phakiti, 2019; Xi, 2010). We chose the argument-based validation (ABV) framework (Aryadoust, 2013; Chapelle, 2020; Chapelle et al., 2008; Kane, 1992, 2006, 2013) to systematize the reviews because validity is the prerequisite of applying AWE systems both in high-stakes assessments and classroom instruction and is the focus of much AWE research conducted so far (Shermis et al., 2013; Warschauer & Ware, 2006). In addition, among the available validity frameworks, we chose the argument-based validation (ABV) framework which deals with the intended interpretations, uses and consequences of assessment scores (Chapelle et al., 2015). Notably, this framework would link automated scoring with construct definition and test design (Bennett & Bejar, 1998) as well as traditional assessment topics like validity and fairness (Mislevy, 2020).

The Argument-based validation framework has gained popularity in research of automated scoring. For example, Xi (2010) proposed a series of research questions for both automated scoring and automated feedback studies based on the ABV framework. In another study, Enright and Quinlan (2010) discussed the use of ABV framework in the validation of human and e-rater scoring in the TOEFL Internet-based Test (TOEFL iBT). In addition, a systematic ABV framework, with research areas, criteria for conducting research on AWE scores and threshold when evaluating automated essay scoring systems, was presented by Williamson, Xi, and Breyer (2012). These evaluating criteria were further explicitly stated and explained as five major aspects: “association with human scores, fairness, relationships with external variables, reliability across tasks and test forms, and impact of use” (Ramineni & Williamson, 2013, p. 32). As discussed below, these aspects can be mapped onto ABV and further expanded.

The Theoretical Framework

The argument-based approach to validity, or argument-based validation (ABV), (Chapelle, 2020; Chapelle et al., 2008; Kane, 1992, 2006, 2013) was developed based on the unitary concept of validity (Messick, 1989) and the Toulmin's informal argumentation model (1958). According to Kane (1992, 2006, 2013), within ABV, there are two interrelated stages: the formative stage that consists of developing an interpretive argument and the summative stage comprising of the development of a validity argument. In the first stage, the purported interpretations and uses of test scores should be clearly stated, and in the second stage, the claims and assumptions made in the first stage are empirically tested. A complete argument, as stated in Toulmin's argument model (2003), is composed of several elements: the data and inference justified by warrants on account of backing, and leading to claims, or claim with qualifier if any rebuttal exists.

A claim is the conclusion or the destination of the argument, and in validity arguments, consists of statements about test scores and their intend impacts (Chapelle, 2020). Data refers to the facts related to and supporting the claim; warrants are the general, hypothetical statements that bridge and authorize the process from the data to the claim (Toulmin, 2003). In addition, backing comprises of current, relevant findings in the form of categorical statements of fact with the purpose of establishing authority for warrants. Qualifiers are used to indicate the strength of the claim, for example, 'certainly' for a strong claim and 'probably' for a weak one, whereas rebuttals are the exceptions or circumstance where the warrant is not applicable, i.e., is attenuated or refuted.

We formulated claims and areas of investigation for each of the inferences of AWE scoring as shown in Appendix A. In line with Chapelle et al. (2008) and Dursun and Li (2021), the ABV framework consists of six inferences: the domain description, evaluation, generalization, explanation, extrapolation, and utilization inferences. Each inference comprises of specific warrants, assumptions, and backing, and may be attenuated by counter-evidence or rebuttals. Each inference and their respective areas of investigation in regards to AWE scoring are discussed next.

The domain description inference is analogous to Messick's (1989) construct definition and links the features of the target language use (TLU) domain to the features of the writing task. Accordingly, the warrant to support this inference consists of evidence that the writing assessment task represents writing tasks in the TLU domain (Chapelle et al., 2008; Xi, 2010). The areas of investigation that would offer backing (or rebuttals) for this inference include domain analysis and task design. Domain analysis refers to the process of gathering information concerning the nature of language in the TLU domain. This is accomplished through the examination of the knowledge and surveys of language and language users in the TLU domain, as well as needs and corpus analysis (Dursun &Li, 2021). Task designs draw upon expert judgments and the empirical analysis of task performance to simulate tasks in the TLU domain.

The evaluation inference in this study links test-takers' performance to the observed scores measured by AWE systems. This inference posits that the observed scores should be accurate representations of test-takers' performance (Weigle, 2013b).

Areas of investigation to collect backing for this inference are human scoring processes and score quality, the agreement of human raters and AWEs, mean score differences between human and automated scoring, and evaluation of automated and human scores at the different levels such as task level and reported score level (Williamson et al., 2012). Human scoring process and score quality serve to guarantee the quality of human scores, which has traditionally been regarded as the gold standard for training, calibrating, and testing AWE systems (Bridgeman, 2013; Williamson et al., 2012). Human-machine agreement analysis investigates the agreement between human raters and AWEs as evidence for reliability, which is widely measured with Cohen's kappa and Pearson correlation.

The generalization inference links the observed scores to the test domain. The warrant for this inference consists of evidence that test-takers' observed scores in one task would be appropriate estimates of expected scores obtained from other similar tasks within the same universe of tasks (Aryadoust, 2013; Chapelle, 2020; Kane, 1992). Areas of investigation include the generalizability of automated scores across parallel versions of tasks and test forms, and the prediction of human scores on alternative/parallel forms. This reliability of automated scores can be examined with generalizability (G), dependency (D), and Phi coefficients (Sawaki & Xi, 2019).

The explanation inference links the expected scores from the previous stage to the construct of interest; that is, the scores are attributed to and represent the construct (Chapelle et al., 2008). This inference is intimately related to the traditional construct validity and, thus, construct validation analyses and investigations fully pertain to it

(Aryadoust, 2013). Areas of investigation include scoring features of AWE systems in relation to the representation of the construct of interest, the scoring rubrics which elucidate the features of the construct (Williamson et al., 2012; Xi, 2010), and the factorial structure and psychometric features of the test scores (Aryadoust, 2013).

The fifth inference, the extrapolation inference, links expected scores to the performance in the TLU domain. Extrapolation is analogous to the traditional criterion-referenced validity, and may therefore be investigated via conventional predictive and concurrent validity analysis (Aryadoust, 2013) as well as analysis of the scoring practice of domain insiders. The warrant consists of evidence that the automated scores are correlated with external indicators of writing ability in the TLU or other related tests or assessments (Williamson et al., 2012; Weigle, 2013a).

The last inference is utilization, which links test scores to their uses and consequences (Chapelle et al., 2008; Enright & Quinlan, 2010). This inference is supported by the evidence that the scores provide meaningful and useful information for making decisions about test takers and designing curricula. Utilization is analogous to traditional washback and/or consequential validity and, accordingly, its backing may be generated via the investigation of the impact of using automated scoring on the accuracy of decisions in high stakes tests made based on test scores (Aryadoust, 2013; Chapelle et al., 2008). In addition, consequences of using automated scoring such as potential changes in users' perceptions, preparation, and teaching should also be examined (Enright & Quinlan, 2010; Xi, 2010). Another area of investigation is the differential impacts of AWEs on different subgroups, which is pertinent to the concept of fairness in assessment (Williamson et al., 2012). It is noted that the issue of fairness may involve more than the inference of utilization; however, we regard this as a component of utilization, as the current ABV frameworks have not explicated fairness as a stand-alone inference.

Based on the outlined framework above, the main aim of this study is to investigate whether, and how, each of the inferences was validated in the previous AWE scoring research. By describing the specific context of these studies, this study also aims to identify the trends of such research and its implications. The specific research questions are as follows:

1. What are the study contexts and the trends of AWE scoring, in terms of locations of studies, target language, language environment, ecological setting, and educational context?

2. Which validity inferences have been investigated in the identified AWE studies?

3. What methodologies were adopted in the AWE studies to investigate each validity inference?

4. What areas of investigation have been studied to provide evidence of backing or rebuttals for each inference?

Research Methodology

Dataset

We used Scopus as the database for the literature search. Scopus has important advantages over other available databases; notably, it is “the largest abstract and citation database of peer-reviewed literature” (Schotten et al., 2018). To avoid missing relevant studies, the search was not limited to any specific journals. We identified the following search items based on the relevant literature (e.g., Stevenson, 2016; Stevenson & Phakiti, 2014): “automated/automatic essay/writing evaluation/evaluator”, “automated/automatic essay/writing scoring/scorer”, “automated/automatic essay/writing assessing/assessment”, “automated/automatic essay/writing rating/rater”, “automated/automatic essay/writing grading”, “automated/automatic essay/writing grading”. We also used the Scopus search code to ensure the comprehensiveness of the research results (see Appendix B for details).

This search yielded a preliminary dataset of 360 papers from Scopus (last retrieved, May 14, 2020). In the meantime, to ascertain that all identified studies were related to our research questions, a further screening process was conducted based on two inclusion and exclusion criteria. The first criterion was that papers should be empirical in nature and published in peer-reviewed journals (Riazi et al., 2018), as empirical studies would be sources of evidence for or against claims in the ABV framework. The second criterion was that the papers were supposed to focus on the AWE scoring. Therefore, papers that were not related to AWE systems, or focused on the feedback by AWE systems were excluded. In the end, six papers were inaccessible and excluded from our dataset. This screening process narrowed down the dataset to 105 papers published in 68 journals[1].

We present the screening process with PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart (Sarkis-Onofre et al., 2021) in Figure 1, which is a conventional way to ensure the transparency and completeness of systematic reviews. The descriptive information of the journals is demonstrated in Appendix C.

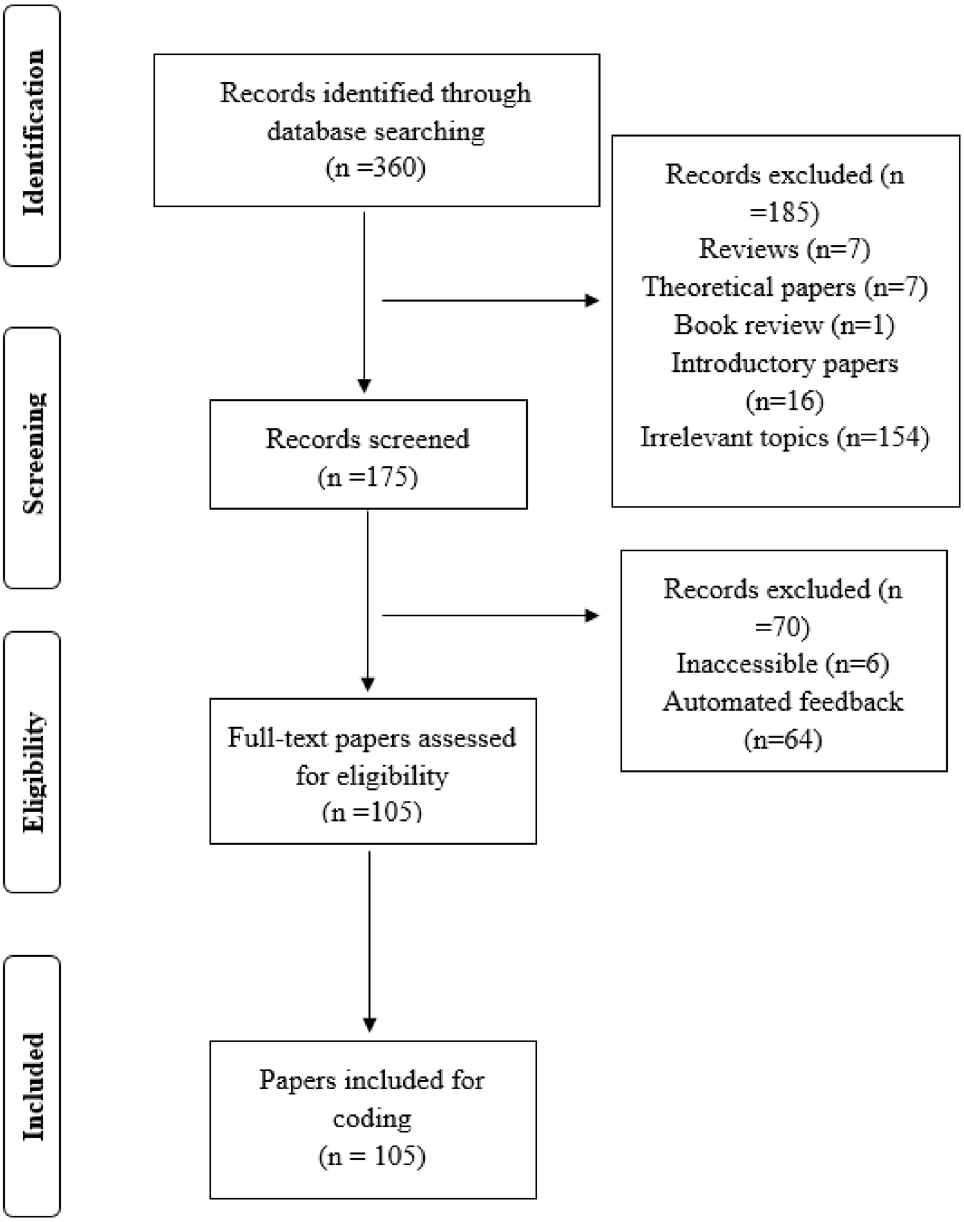
Identification
Screening
Eligibility
Included
Records identified through database searching (n =360)
Records excluded (n =185) Reviews (n=7) Theoretical papers (n=7) Book review (n=1) Introductory papers (n=16) Irrelevant topics (n=154)
Records screened (n =175)
Records excluded (n =70) Inaccessible (n=6) Automated feedback (n=64)
Full-text papers assessed for eligibility (n =105)
Papers included for coding (n = 105)

Figure 1 PRISMA flowchart of data screening

Coding Scheme

A coding scheme was designed based on studies by Dursun and Li (2021), Fan and Yan (2020), Stevenson and Phakiti (2014), and Zhang and Yu (2019) (see Appendix D). The unit of coding was set to be the "research study", which was defined as a set of data that was collected and investigated under a single research plan from a sample of participants (Stevenson & Phakiti, 2014). In the 105 papers, 5 publications included two research studies, each with different samples, resulting in 110 studies in total for coding and analysis.

All variables in the coding scheme were grouped into five categories that correspond to the foregoing research questions: administrative information, study context, inferences of validity, research design, and results. Administrative information refers to general information of the publications and consists of authors, titles, years, and source of journals which provide descriptive statistics of the studies involved. The study context was defined in terms of study location, language environment, target language, ecological setting (e.g., classroom setting & high-stakes tests) (Shermis et al., 2013), and educational context. Specific definitions for each category within the variables are presented in appendix D. In addition, the six inferences (i.e., domain description, evaluation, generalization, explanation, extrapolation, and utilization) listed in Appendix A were used to code each study. As most studies did not adopt the ABV framework, we coded the inferences based on a close examination of the abstract and research questions of each study.

The research design consists of research methodology ‘type 1’ and ‘type 2’. These two research types were used to code methodologies from different angles: (i) the first one was adopted from Riazi et al. (2018) and consisted of four categories: qualitative, quantitative, eclectic, and mixed methods; and (ii) type 2 implemented a binary code to record the study from a temporal perspective, thereby grouping the studies into cross-sectional and longitudinal studies (Phakiti et al., 2018).

The category of results subsumed four categories: results of backing, results of rebuttals, areas of investigation for backing, and areas of investigation for rebuttals. For results of backing and rebuttals, we relied on the relevant findings or results as stated in the abstract and/or the results section in the paper. For areas of investigation, we coded each study into one or more areas (see Appendix A) as discussed in the ABV framework based on research questions and results of each study.

Reliability

To ensure reliability, both authors examined the coverage of the dataset, and the processes of data generation from Scopus together. The design of the coding scheme was done by both authors after five rounds of discussion; and the coding process underwent two iterative stages to ensure reliability: first, the first author coded all articles twice with an interval of three months, next, all data was reviewed and checked by the second author. The intra-coder reliability was calculated with Cohen’s Kappa, and the kappa’s value for each variable was above 0.90.

Results

Research Question 1

Table 1 presents the study contexts of AWE scoring research. Regarding the target language, in total, 11 languages were examined in the dataset; nevertheless, English was the most frequently studied language (n = 95, 86.37%), while languages other than English were rarely investigated. The second variable is the language environment. Six categories were identified: first language, foreign language, second language, a combination of first language and second language, a combination of foreign language and second language, and a mix of first, second, and foreign language. The highest number of studies was conducted in first language contexts (n=51, 46.37%). For ecological setting, the number of studies conducted in classrooms was the largest among all study environments (n=63, 57.27%), followed by high stakes assessment (n=36; 32.73%). Lastly, regarding educational level, the fourth variable, almost half of the research was conducted in university contexts.

Figure 2 presents the trend of AWE scoring research. Overall, a generally rising interest in AWE scoring was noticed. The increase became more noticeable around 2011, as increasingly more research was carried out to investigate AWE scoring.

Table 1. *Study Context of AWE scoring*

Target language	No. of papers	Language Environment	No. of papers	Ecological setting	No. of papers	Educational level	No. of papers
English	95	L1	51	classroom	63	University	52
Chinese	3	FL	21	assessment	36	K-12	31
Arabic	3	L1+L2	13	online course	5	unspecified	21
French	2	Mixed	11	unspecified	6	K-12 & University	5
Finnish	2	L2	5			mixed	1
Spanish	1	unspecified	5				
Punjabi	1	FL+L2	4				
Portuguese	1						
Indonesian	1						
Hebrew	1						
German	1						

1. FL= foreign language; L1= first language; L2=second language; L1+L2= a combination of first language and second language; Mix=a combination of first, second and foreign language; L2+FL= second language and foreign language. See Appendix D for their explanation

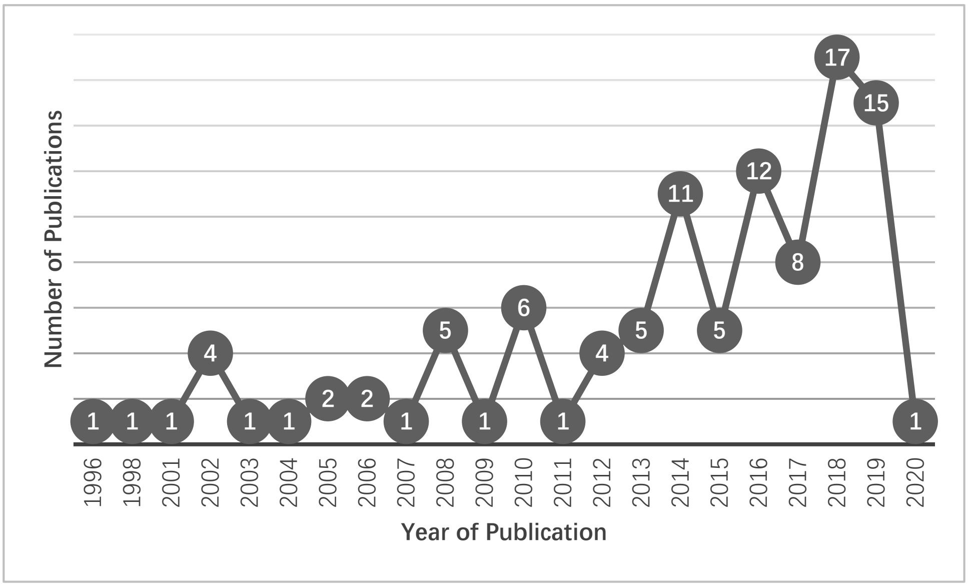

Figure 2 Trend of AWE scoring publications over the years.

Research Question 2

Table 2 shows research studies investigating the six validity inferences of AWE scoring within the ABV framework. It needs to be pointed out that too few studies clearly stated using ABV framework (n = 4), or traditional validities (n = 12). Therefore, we coded the studies based on their research questions and results and their link with the validity inferences in ABV. Our results showed that while all six inferences were investigated in these studies, there was an uneven distribution of evidence for the inferences across the studies. Among the studies, the evaluation inferences are the most studied type (n=86, 78.18%). The generalization, explanation, and extrapolation inferences were similar to each other in terms of the number of studies conducted. Compared to the above inferences, the domain description inference seemed to be neglected, with only two studies being done.

Table 2. *Representation of Six Validity Inferences in the Dataset (n=110)*

Inferences	# of studies	%
Domain description	2	1.82
Evaluation	86	78.18

Generalization	13	11.82
Explanation	15	13.64
Extrapolation	14	12.73
Utilization	25	22.73

Notes:

1. Four studies explicitly adopted ABV, while 12 other studies explicitly adopted a traditional view of validity; the rest of the studies did not explicitly mention validity.

2. Seventy-six studies examined one inference each; 26 studies examined two inferences; 6 studies examined three inferences each; one study examined four inferences, and one study five inferences.

Research Question 3

Table 3 displays the summary statistics of methodologies for the validity inferences under investigation. Of all the studies, an overwhelming majority adopted quantitative methods (n = 99, 90%), with only two studies adopting qualitative methods. The number of studies using explicit mixed methods (n = 5) was the same as the number of studies implicitly adopting mixed methods (coded as Eclectic). In addition, significantly more studies (n = 92, 83.63%) were cross-sectional in nature, compared to longitudinal studies (n = 18, 16.36%).

The methodologies used for each inference are also vast and varied. Two studies examining the domain description inference adopted quantitative and qualitative methods separately. For the evaluation inference, most of the studies were conducted quantitatively (n = 78, 70.90%) as compared to studies using other methods, while only one study adopted the qualitative method to approach this inference. At the same time, most of the studies were also cross-sectional when examining this inference. Similar patterns were identified for the generalization, explanation, and extrapolation inferences: almost all the studies on these inferences (n = 13; n = 14; n = 13) used quantitative methods, and none applied qualitative analysis. Only one study used an eclectic method to study the explanation inference, and another one adopted mixed methods to study the extrapolation method. With regards to the utilization inference, the largest number of the studies also used quantitative methods (n = 20).

Table 3. *Methodologies for Six Validity Inferences in the Studies*

Methodology	Domain description		Evaluation		Generalization		Explanation		Extrapolation		Utilization	
	# of studies	%	# of studies	%	# of studies	%	# of studies	%	# of studies	%	# of studies	%
Quantitative (n=99)	1	0.91	78	70.90	13	11.82	14	12.73	13	11.82	20	18.18
Qualitative (n=1)	0	0	1	0.91	0	0	0	0	0	0	0	0
Mixed methods (n=5)	1	0.91	3	2.73	0	0	0	0	1	0.91	2	1.82
Eclectic (n=5)	0	0	4	3.64		0	1	0.91	0	0	3	2.73
Cross-sectional (n=92)	2	1.82	74	67.27	10	2.73	14	12.73	8	7.27	15	13.64
Longitudinal (n=18)	0	0	12	10.91	3	10.91	1	0.91	6	5.45	10	9.09

Research Question 4

Tables 4 present the results of areas of investigation in the backing and rebuttals of each validity inference. Overall, for each inference, evidence supporting and weakening the inferences was found except for the domain description inference, with the ratio of studies offering backing for each inference being much higher.

For the domain description inference: both studies were reported in one paper (i.e., Burstein et al., 2016), and investigated the needed genres of writing in the postsecondary education context, such as annotated bibliographies as well as research proposals in addition to conventional essays with the aim of informing AWE design and development. No statements about rebuttals were identified in these two studies though the researcher found a mismatch among required genres of writing in secondary school, post-secondary institutions, and workplaces.

Regarding the evaluation inference, of all the 86 studies, most yielded supporting evidence, thus providing backing for this inference. The backing evidence concentrated on human-machine agreement; that is, these studies reported satisfactory correlations between automated scores and human scores, but it should be noted that the criteria were not unified. Human scoring processes and score quality was also explored in three studies, which provided evidence backing the automated scoring process. On the other hand, there were still 21 studies yielding rebuttals in the area of human-machine agreements, reporting unsatisfactory or low agreement between human scores and automated scores. Two studies also found evidence of standardized mean score differences between human-yielded and automated scores: Ramineni and Williamson (2018) found that the AWE system, e-rater, was not severe enough in penalizing certain language errors, while Liu and Kunnan (2016) reported that the AWE system, WrtieToLearn, was more stringent in scoring than human raters.

In the studies of the generalization inference (n = 13), 11 studies generated results to support the generalizability of automated scores across parallel tasks or raters; that is, no significant differences were found across different prompts or different genres. Three studies (Klobucar et al., 2013; Qian et al., 2020; Vajjala, 2018), on the other hand, noted situations where little correlation or significant differences were found in automated scoring across different writing tasks, providing rebuttals for the inferences.

The explanation inference and the extrapolation inference shared a similar pattern with the generalization inference, both in the total number of studies, and in terms of the number of studies that yielded supporting evidence or attenuating results (i.e., rebuttals). The studies that reported evidence relating to the explanation inference (n = 15) focused on the scoring or assessing features in relation to the representation of the constructs of interest. Among these, 13 studies provided evidence to show that AWE scoring measures the writing construct in ways human raters do. Still, two studies found counterevidence weakening this inference such as lack of recognition for sentence variety or structural aspects of essays (Lee et al., 2010) and significant bias towards word counts (Perelman, 2014). In addition, 14 studies examined the extrapolation inference of AWE scoring. Of these, 13 studies investigated the topics that are aligned with the extrapolation inference by examining the relationship between automated scores and external measures of writing construct and identified significant correlations with other indicators of writing ability (e.g., Attali, 2015; Bridgeman & Ramineni, 2017). Two studies (i.e., Powers et al., 2002b; Reilly et al., 2014) found that such correlations were weaker and lower compared to the correlation between human scores and external measures.

In the utilization inference, a higher percentage of counterevidence was found, undermining this inference. This first area of investigation in this inference was the consequence of using automated scoring on writing, wherein 10 studies provided positive results (e.g., Grimes & Warschauer, 2010), while in three studies (i.e., Gerard & Linn, 2016; Tsai, 2012; Wilson, 2017), little improvement in students' writing performance were also reported. Additionally, in five studies (e.g., Li et al., 2015), teachers and/or students viewed AWE scores more negatively compared to the scores provided by human raters. Another area of investigation under this inference was fairness of scores for various subgroups, in which the results were mixed. Three studies (i.e., Bridgeman & Ramineni, 2017; James, 2008; Ramineni & Williamson, 2018) showed mixed results: that is, the fairness was achieved for some subgroups as shown by low discrepancy between automated scores and human scores for different subgroups, but not for other subgroups. The impact of using automated scoring on the accuracy of decisions is also one area of investigation for this inference; three studies (i.e., Bridgeman & Ramineni, 2017; James, 2008; Reilly et al., 2016) examined this area, all of which yielded supporting results for this inference.

Table 4. *Areas of Investigation for Each Inference*

Areas of investigation	Backing		Rebuttal	
	# of studies	Publication Numbers[1]	# of studies	Publication Numbers
Domain description (n=2)				
domain analysis	2	10,11	0	N/A
Evaluation (n=86)				
Human–machine agreement	71	1, 2, 4, 5, 7, 12, 14, 15, 16, 17, 19, 20, 22, 24, 25, 28, 29, 30, 32, 33, 34, 35, 36, 37, 41, 42, 43, 44, 45, 46, 47, 49, 50, 52, 53, 54, 55, 58, 59, 60, 61, 62, 63, 64, 66, 68, 73, 74, 76, 77, 79, 80, 82, 83, 85, 86, 87, 89, 90, 92, 94, 95, 97, 103, 104, 105, 106, 107, 108, 109, 110,	21	15, 26, 27, 31, 35, 49, 51, 56, 57, 67, 70, 75, 76, 77, 79, 91, 92, 93, 96, 97, 103,
Human scoring process and score quality	4	48, 69, 72,103	0	N/A
Standardized mean score differences between human scoring and automated scoring	0	N/A	2	54, 75

Areas of investigation	Backing		Rebuttal	
	# of studies	Publication Numbers	# of studies	Publication Numbers
Generalization (n=13)				
Generalizability of scores across parallel versions of tasks / test forms / across raters	11	4, 28, 36, 37, 38, 42, 46, 66, 74, 83, 100	3	37, 71, 94
Explanation (n=15)				
Scoring features vis- à-vis the representation of construct of interest	13	14, 18, 37, 40,48, 56, 58, 61,66, 69,88, 94, 102	3	8, 48, 65
Extrapolation (n=14)				
External relationships	13	4, 5, 9, 17, 37, 39, 46, 60, 68, 74, 83, 97, 102	2	68, 76
Utilization (n=25)	25			
Consequences of using automated scoring	10	3, 6, 21, 23, 37, 39, 51, 80, 91, 99,	3	21, 92, 99
Fairness for subgroups	6	9, 12, 28, 74, 75, 106	4	9, 28, 75, 78
Impact of using automated scoring on the accuracy of decisions made based on test scores	3	29, 98, 101	0	N/A
Teachers/students' perception of automated scoring	1	81	5	13, 23, 31, 51, 84

Note: 1. All publications are listed in the supplementary file. 2. N/A = not applicable

Discussion

In this study, we systematically reviewed the AWE scoring studies that were indexed in Scopus. The AWE scoring papers spanned across multiple disciplines as indicated by the journals where the studies were published. To synthesize the validity evidence from these studies, we used the ABV framework (Aryadoust, 2013; Chapelle, 2020; Chapelle et al., 2008; Kane, 1992, 2006, 2013) and coded the study context, methodologies, inferences, and results to shed fresh light on research in and application of AWE scoring. The research questions of the study are discussed next.

Research Question 1

Research question 1 was designed to investigate the study contexts and the trends of AWE scoring research. It was found that the included AWE scoring studies were highly heterogeneous, varying in terms of the systems being studied, the target language, the language environment, the ecological settings, and the educational level.

What deserves attention, though not explicitly discussed in these studies, was the potential influence of the heterogeneous contexts on the construct of writing. Possible relevant questions might include, for example: will the writing construct be different from first language environment to foreign language environment, from classroom assessment to high-stakes tests, or from high school to university, or even from one AWE scoring system to another system? For example, Weigle (2013b) made a clear distinction between English language learners (ELL) in English-medium education systems like the US and EFL learners in their own countries and conceptualized the writing construct for ELLs based on this difference. In other words, it is essential to determine whether the reliability and validity of the AWE scoring systems developed for L1 and/or L2 evaluation is consistent across different contexts. In addition, recent research by Lamprianou et al. (2020) found that human raters' longitudinal performance changes significantly over time. Therefore, it is necessary to examine the longitudinal performance of the AWE scoring systems used in high-stakes decision-making and whether there are any sources of bias affecting different cohorts of test-takers with different L1s and L2s.

Research Question 2

Research question 2 was set to investigate the representation of validity inferences in the dataset. The results showed an unbalanced distribution of studies pertinent to each validity inference

The most studied inference is the evaluation inference, which was examined in over two-thirds of the papers (n = 86, 78.18%). However, the finding is not surprising as the thrust of AWE research, according to Stevenson and Phakiti (2014) has been validating AWE scores using psychometric and statistical methods such as measuring the correlation between automated scores and human scores. On the other hand, a smaller number of studies were conducted on the generalization, explanation, and extrapolation inferences when validating AWE scoring. These three inferences respectively involve reliability, construct validity, and the defining of reference criteria in validity research (criterion-referenced validity) (Aryadoust, 2013), which have been recognized as the most important evidence of validity since, at least, the publication of the seminal validity paper by Cronbach and Meehl (1955). As AWE scoring systems were initially designed to score writing in high-stakes assessments (Shermis et al., 2013), these inferences should be validated with more rigorous methods, especially for the newly developed AWE scoring systems. Accordingly, we suggest that, rather than focus solely on examining and improving the agreement between human scoring and AWE scoring, research efforts should also be directed to investigating these three validity inferences. Particularly, the concern of opposing scholars should be underscored: that many, if not all, AWE systems do not capture deep layers of discourse in writing assessments. Recent developments in deep learning and AI would provide scholars with an avenue for interdisciplinary research into AWE systems to enhance the weaker inferences in the validity argument of these systems.

Finally, the domain description inference was rarely studied; however, the importance of this inference is evident in that it is the starting point of test development and argument-based validity (Chapelle et al., 2008). That is, domain description pertains to the delineation of the TLU domain and the construct under assessment. Im et al. (2019) argued that domain analysis is “the most important aspect for test design to identify language knowledge, skills, and abilities…” (p. 19). Likely due to this gap in knowledge, the construct coverage by AWE scoring systems has often been a source of criticism (Deane, 2013). Though it is possible that this inference was taken into consideration before or while designing the AWE scoring programs, many AWE scoring systems, particularly those developed by researchers themselves, just borrowed training data from public corpora such as essay sets by Automated Student Assessment Prize (ASAP) (Shermis, 2014). As such, it is necessary to examine the domain description inference before moving on to the higher-level validity inferences. Notably, distinctive features of the domain for which AWEs are developed should be identified and built into construct definition statements. The corpora used to train algorithms should also be chosen carefully to represent the tasks and discourse of the target language use (TLU) domains. That is, using general corpora like those used in ASAP for a TLU situation with minimal resemblance to the ASAP's potential TLU domain would impinge on construct definition and operationalization, particularly if these TLU domains have diverging features.

Research Questions 3 and 4

Research questions 3 and 4 were set to investigate the methodologies and results for each validity inference. Overall, most studies adopted quantitative methodologies and yielded positive evidence, thereby providing various degrees of backing for each inference.

We found that quantitative methods were dominant in the studies reviewed. In these studies, supporting validity evidence consisted of the percentage of exact and adjacent agreement, Cohen's Kappa, Pearson's correlation coefficient, standardized mean difference, and so forth (Rotou & Rupp, 2020; Williamson et al., 2012). In terms of validation, although the ABV framework was proposed for validating AWE scoring systems by many researchers (e.g., Weigle 2013a, b; Williamson et al., 2012; Xi, 2010), it was surprising to find that only four studies explicitly adopted this framework, while most research focused on conventional ways of evaluating the human and machine scores agreement, which might be due to the large number of various systems being examined in these studies.

The strength of ABV framework lies in its chain of inferences (i.e., six inferences) in validating the interpretation and uses of test scores (Kane, 2004, 2013). Any lower-level conclusion/claim functions as a premise and data for the higher-level inference, and if the former were not validated, it would be unjustified to jump to the next inference and argue for the validation of the next. As such, the domain description inference, which was studied only in one paper, was improperly neglected based on its fundamental role as the first inference in the ABV framework. This would have a negative effect propagated upwards into the higher inferences in the ABV framework. In this sense, the chain of inferences within the ABV framework helps us point out directions for needed research by stating the claims, warrants, and importantly, evidence they would require. We concede that how the attenuation of a lower-level inference would affect the higher-level inference is underresearched in language assessment. This can be due to the nature of ABV studies which are often post-hoc; that is, they are developed after the assessments are created and fully rolled out. We suggest future research should address this gap in AWE scoring systems.

Moreover, according to Kane (2013), the validity of the proposed score interpretations and uses depends on how well the evidence could support such claim(s), and more ambitious claims require more supporting evidence. Therefore, if AWE scoring systems are to be used in higher stakes tests, more backing evidence needs to be collected and counterevidence, if any, should be further investigated. In this sense, the studies yielded convincing evidence supporting the use of quite a few AWE scoring systems, while the key validity inferences remain relatively less researched. This paucity of evidence along with the rebuttals for some of the inferences warrant further investigation. For example, Perelman (2014) noticed the huge bias of AWE scoring systems towards word count which might encourage response strategies that result in construct-irrelevant variance and undermine the explanation inference. When deciding on the use of similar systems, researchers need to reexamine this inference for validation.

Conclusion

Our review was an attempt at providing a comprehensive and systematic review of AWE scoring research by examining AWE-related literature indexed in Scopus within the argument-based validation framework. Positive results were found in these studies, which supports the inferences within the ABV framework and suggests that AWE scoring could be used in high-stakes tests and in classrooms, or online courses. Nevertheless, what needs to be cautioned is the way of using these systems. For now, to completely replace human raters with AWE scoring system is unjustified as counterevidence was identified for each validity inference; instead, AWE scoring might better be used in combination with human scoring in high-stakes tests or as an independent rater in formative assessments that can provide immediate scores to language learners.

Finally, as suggested by Richardson and Clesham (2021), to ensure validity, the use of AI technology in assessment must be domain- and context-dependent. As we showed, most of the AWE research, however, failed to explicitly tap into the TLU domain or the domain description inference. As this inference is the starting point of the argument-based validity and serves as the premise of higher-order inferences (Chapelle, 2020), we suggest that, in future research, a research-based (rather than opinion-based) definition of the TLU domains needs to be presented at this stage of assessment development. This will allow researchers to base the definition and specification of the attributes of the TLU on a more objective set of criteria which will also inform test or AWE scoring program developers.

Acknowledgement

XXX

Notes:

1. The papers included are numbered and listed in a supplementary file which can be found from the following link: XXX.

References

Aryadoust, V. (2013). Building a Validity Argument for a Listening Test of Academic Proficiency. Cambridge Scholars Publishing

*Attali, Y. (2015). Reliability-Based Feature Weighting for Automated Essay Scoring [Article]. *Applied Psychological Measurement*, 39(4), 303-313. https://doi.org/10.1177/0146621614561630

Bennett, R. E., & Bejar, I. I. (1998). Validity and Automated Scoring: It's Not Only the Scoring. *Educational Measurement: Issues and Practice*, 17(4), 9–17. https://doi.org/10.1111/j.1745-3992.1998.tb00631.x

Brew, C., & Leacock, C. (2013). Automated short answer scoring: Principles and prospects. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. (pp. 158–174). Routledge/Taylor & Francis Group.

*Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings [Article]. *Assessing Writing*, *34*, 62-71. https://doi.org/10.1016/j.asw.2017.10.001

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco: Mexico

*Burstein, J., Elliot, N., & Molloy, H. (2016). Informing automated writing evaluation using the lens of genre: Two studies [Article]. *CALICO Journal*, *33*(1), 117-141. https://doi.org/10.1558/cj.v33i1.26374

Burstein, J., Riordan, B., & McCaffrey, D. (2020). Expanding automated writing evaluation In Yan, D., Rupp, A. A., & Foltz, P. (Eds). *Handbook of Automated Scoring: Theory into Practice.* (pp. 329–346). New York, NY: Taylor and Francis Group/CRC Press

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Chapelle, C. A. (2020). *Argument-Based Validation in Testing and Assessment (Quantitative Applications in the Social Sciences)* (1st ed.). SAGE Publications, Inc.

Chapelle, C., Enright, M., & Jamieson, J. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. Routledge.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385–405. https://doi.org/10.1177/0265532214565386

Chapelle, C. A., & Voss, E. (2021). *Validity Argument in Language Testing: Case Studies of Validation Research (Cambridge Applied Linguistics).* Cambridge University Press.

*Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education,* 31(3), 241–250. https://doi.org/10.1080/08957347.2018.1464450

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. https://doi.org/10.1037/h0040957

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002.

Dursun, A. & Li, Z. (2021) A systematic review of argument-based validation studies in the field of Language Testing (2000–2018). C. Chapelle & E. Voss (Eds.), *Validity Argument in Language Testing: Case Studies of Validation Research (Cambridge Applied Linguistics)*. (pp. 45-70) Cambridge University Press.

Ericsson, P. F. & Haswell, R. (Eds.). (2006). *Machine Scoring of Student Essays: Truth and Consequences.* Utah State University Press.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring [Article]. *Language Testing*, *27*(3), 317-334. https://doi.org/10.1177/0265532210363144.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11,* 330. https://doi.org/10.3389/fpsyg.2020.00330

*Gerard, L. F., & Linn, M. C. (2016). Using Automated Scores of Student Essays to Support Teacher Guidance in Classroom Inquiry. *Journal of Science Teacher Education*, 27(1), 111-129. https://doi.org/10.1007/s10972-016-9455-6

*Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, 8(6).

Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia*, 9(1), 14.

*James, C. L. (2008). Electronic scoring of essays: Does topic matter? *Assessing Writing*, 13(2), 80-92. https://doi.org/10.1016/j.asw.2008.05.001

Jang, E.E., & Wagner, M. (2013). Diagnostic feedback in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 693-711). Wiley.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)*, (pp. 17-64). American Council on Education, Praeger Series on Higher Education.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. (pp. 147- 168). Erlbaum.

*Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated Scoring in Context: Rapid Assessment for Placed Students. *Assessing Writing*, 18(1), 62–84. https://doi.org/10.1016/j.asw.2012.10.001

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2020). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing*. https://doi.org/10.1177/0265532220940960

Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st International Workshop on Gender Equality in Software Engineering (pp. 14-16). https://doi.org/10.1145/3195570.3195580

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*, 3, 1–34. https://doi.org/10.2200/S00275ED1V01Y201006HLT009

*Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, 27, 1-18. https://doi.org/10.1016/j.jslw.2014.10.004

Li, S., & Wang, H. (2018). Traditional literature review and research synthesis. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.) *The Palgrave Handbook of Applied Linguistics Research Methodology* (pp. 123-144). Palgrave-MacMillan.

*Matthews, J., & Wijeyewardene, I. (2018). Exploring relationships between automated and human evaluations of L2 texts. *Language Learning and Technology*, 22(3), 143-158.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test, *Language Assessment Quarterly*, (8)2, 161-178. https://doi.org/10.1080/15434303.2011.565438

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Mislevy, R. (2020). An evidentiary-reasoning perspective on automated scoring: Commentary on part I In Yan, D., Rupp, A. A., & Foltz, P. (Eds). *Handbook of Automated Scoring: Theory into Practice*. (pp. 151–167). Taylor and Francis Group/CRC Press

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*, 6(7), e1000097. https://doi.org/10.1371/journal.pmed1000097

National Council of Teachers of English. (2013, April). *NCTE position statement on machine scoring.* https://ncte.org/statement/machine_scoring/

*Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, 21, 104-111. https://doi.org/10.1016/j.asw.2014.05.001

Perin, D., & Lauterbach, M. (2018). Assessing Text-Based Writing of Low-Skilled College Students [Article]. *International Journal of Artificial Intelligence in Education*, 28(1), 56-78. https://doi.org/10.1007/s40593-016-0122-z

*Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002a). Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2), 103–134. https://doi.org/10.1016/s0747-5632(01)00052-8

*Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002b). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research,* 26(4), 407-425. https://doi.org/10.1092/UP3H-M3TE-Q290-QJ2T

*Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's Automated Essay Scoring System iWrite [Article]. *Journal of Educational Computing Research*, 58(4), 771-790. https://doi.org/10.1177/0735633119881472

*Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® General Test. *ETS Research Report Series*, 2018(1). doi:10.1002/ets2.12192

*Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distance Learning*, 15(5), 83-98. https://doi.org/10.19173/irrodl.v15i5.1857

Riazi, M., Shi, L., & Haggerty, J. (2018). Analysis of the empirical research in the journal of second language writing at its 25th year (1992–2016). *Journal of Second Language Writing*, 41, 41–54. https://doi.org/10.1016/j.jslw.2018.07.002

Richardson, M. & Clesham, R. (2021) 'Rise of the machines? The evolving role of AI technologies in high-stakes assessment'. *London Review of Education*, 19 (1), 9, 1–13. https://doi.org/10.14324/LRE.19.1.09

*Rupp, A., Casabianca, J., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: a case study in Switzerland and Germany. *ETS Research Report Series*, 2019(1), 1–23. https://doi.org/10.1002/ets2.12249

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., & Lockwood, C. (2021). How to properly use the PRISMA Statement. *Systematic Reviews*, 10(1).

Schotten, M., Aisati, M., Meester, W. J. N., Steigninga, S., & Ross, C.A. (2018). A brief history of Scopus: The world's largest abstract and citation database of scientific literature. In F. J. Cantu-Ortiz (Ed). *Research analytics: Boosting university productivity and competitiveness through Scientometrics* (pp. 33-57). Taylor & Francis.

*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). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. xiii–xvi). Lawrence Erlbaum Associates.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. (pp. 1–15). Routledge/Taylor & Francis Group.

Shermis, M., Burstein, J., Elliot, N., Miel, S., & Foltz, P. (2016). Automated writing evaluation: A growing body of knowledge. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research.* (pp. 395–409). Guilford Press.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing,* 19, 51-65. doi:10.1016/j.asw.2013.11.007

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing: Contexts and Issues* (pp. 125-142). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108635547.009

Sawaki, Y., & Xi, X. (2019). Univariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment* (Vol. 1). (pp.30-53). Routledge

Toulmin, S. E. (2003). *The Uses of Argument (Updated ed.)*. Cambridge University Press.

*Tsai, M. H. (2012). The Consistency Between Human Raters and an Automated Essay Scoring System in Grading High School Students' English Writing. *Action in Teacher Education*, 34(4), 328-335. https://doi.org/10.1080/01626620.2012.717033

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, 28(2), 97-111.

*Vajjala, S. (2018). Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features [Article]. International Journal of Artificial Intelligence in Education, 28(1), 79-105. *https://doi.org/10.1007/s40593-017-0142-3*

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, 45(4), 769-774. https://doi.org/10.5054/tq.2011.272525

Warschauer, M., & Grimes, D. (2008). Automated Writing Assessment in the Classroom. *Pedagogies*, 3(1), 22-36.

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

Weigle, S. C. (2013a). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. (pp. 36–54). Routledge/Taylor & Francis Group.

Weigle, S. C. (2013b). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1), 85–99.

*Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, 30(4), 691-718. https://doi.org/10.1007/s11145-016-9695-z

Williamson, D., Xi, X., & Breyer, F. (2012). A Framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27 (3), 291–300

Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from Assessing Writing (2000–2018). *Assessing Writing*, 42. https://doi.org/10.1016/j.asw.2019.100421

Note: * refers to papers that are also included in the dataset.