CJLT RCAT
Canadian Journal of Learning and Technology
La Revue canadienne de l'apprentissage et de la technologie

Volume 48 (3) Fall /Automne 2022

Automated Scoring of Speaking and Writing: Starting to Hit its Stride

Notation automatisée de l'expression orale et écrite : Un début prometteur

***Daniel Marc Jones**, Queen's University at Kingston*

***Liying Cheng**, Queen's University at Kingston*

***M. Gregory Tweedie**, University of Calgary*

Abstract

This article reviews recent literature (2011–present) on the automated scoring (AS) of writing and speaking. Its purpose is to first survey the current research on automated scoring of language, then highlight how automated scoring impacts the present and future of assessment, teaching, and learning. The article begins by outlining the general background of AS issues in language assessment and testing. It then positions AS research with respect to technological advancements. Section two details the literature review search process and criteria for article inclusion. In section three, the three main themes emerging from the review are presented: automated scoring design considerations, the role of humans and artificial intelligence, and the accuracy of automated scoring with different groups. Two tables show how specific articles contributed to each of the themes. Following this, each of the three themes is presented in further detail, with a sequential focus on writing, speaking, and a short summary. Section four addresses AS implementation with respect to current assessment, teaching, and learning issues. Section five considers future research possibilities related to both the research and current uses of AS, with implications for the Canadian context in terms of the next steps for automated scoring.

Keywords: Automated scoring of language; Literature review; Scoring feedback; Technology in language assessment and teaching

Résumé

Cet article examine la littérature récente (2011 jusqu'à présent) sur la notation automatisée (NA) de l'expression écrite et de l'expression orale. Son objectif est d'abord d'examiner les recherches actuelles sur la notation automatisée de la langue, puis de mettre en évidence l'impact de la notation automatisée sur le présent et l'avenir de l'évaluation, de l'enseignement et de l'apprentissage. L'article commence par décrire le contexte général des problèmes de notation automatisée dans l'évaluation et

les tests linguistiques. Il positionne ensuite la recherche sur la NA par rapport aux avancées technologiques. La deuxième section décrit en détail le processus de recherche de la revue de la littérature et les critères d'inclusion des articles. Dans la troisième section, les trois principaux thèmes qui se dégagent de l'analyse sont présentés : considérations relatives à la conception de la notation automatisée; le rôle des humains et de l'intelligence artificielle; et la précision de la notation automatisée avec différents groupes. Deux tableaux montrent comment des articles spécifiques ont contribué à chacun des thèmes. Ensuite, chacun des trois thèmes est présenté plus en détail, avec un accent séquentiel sur l'expression écrite, l'expression orale et un bref résumé. La quatrième section aborde la mise en œuvre des NA en ce qui concerne les questions actuelles d'évaluation, d'enseignement et d'apprentissage. La cinquième section présente les possibilités de recherche futures liées à la recherche et aux utilisations actuelles de la NA, avec des implications sur le contexte canadien en ce qui concerne les prochaines étapes de la NA.

Mots-clés : Notation automatisée de la langue ; revue de littérature ; rétroaction sur la notation ; technologie dans l'évaluation et enseignement des langues

Background on Automated Scoring

Automated scoring (AS) has been the focus of ongoing academic research and development since the 1960s and has enjoyed increasing attention alongside technological advances (Foltz et al., 2020). In the context of language assessment, AS can be defined as "…using computers to convert students' performance on educational tasks into characterizations of the quality of performance" (Foltz et al., 2020, p. 1). The reasons behind these efforts are varied but often relate to cost reduction, scalability, and capacity for immediate feedback or results, as well as consistency and accuracy of assessment (Foltz et al., 2020). Leveraging these benefits through advances in computing, natural language processing, and machine learning, as well as accessible and cost-efficient technological applications has potential advantages for language assessment, teaching, and learning.

Given the critical role of language in facilitating or limiting opportunities for collaboration and employment access in our connected world (McNamara, 2005; Sackett et al., 2001; Shohamy, 2013), as well as the current systemic challenges of mobilizing assessment, teaching, and learning during a global pandemic (d'Orville, 2020; Voogt & Knezek, 2021), AS is poised to play an important role. Increasing AS implementation can increase access to language assessment, as well as teaching and learning resources, while also deepening our understanding of how to optimize their mobilization and implementation—which the field of language assessment is still grappling with (Schneider & Boyer, 2020). The rapid pace of shifting technological affordances (Wood, 2020) offers both new promise and challenges as many of these technologies require additional application and research to ensure validity and fairness. Facilitating valid, reliable, fair, and easy-to-access AS support of language assessment, teaching, and learning based on a wide range of purposes can help alleviate some of the most critical educational challenges faced today. Some of these educational challenges are assessment bottlenecks for certification or training of internationally trained professionals; limited access due to socio-

economic, pandemic-related, or location-specific pressures; and costs related to developing, operationalizing, and maintaining language scoring approaches.

This article provides a big-picture perspective by synthesizing established research, clarifying current AS capabilities, and pointing to future AS directions. Thus, it informs interested language assessment stakeholders by drawing together current research literature on AS. In essence, this is a scoping review of the literature on state-of-the-art language assessment research. Generally, the four language skills considered are speaking, listening, reading, and writing. Listening and reading are receptive language skills, whereas speaking and writing are productive language skills (Golkova & Hubackova, 2014). Though all four language skills have received attention related to AS, writing and speaking have been the subject of more research and operational implementation (Cahill & Evanini, 2020). Given this established research footing and the clear relevance of productive language skills for education, social participation, and workforce mobilization, the focus of this review is on writing and speaking.

Criteria for Selecting Studies

Six major databases were selected for the search: Canadian Business and Current Affairs, Education Source, ERIC, PsychNet, Web of Science, and Academic Search Complete. Articles published before 2011 were excluded to ensure the search findings represented state-of-the-art technologies and up-to-date application of AS. Search items were filtered to include empirical articles, technical reports, and literature reviews written in English. The three primary search terms used in the search were "language assessment," "language evaluation," and "language testing." Each database was searched using this primary search term in combination with each of the following nine secondary search terms. The nine secondary search terms were "artificial intelligence," "natural language processing," "deep neural networks," "machine learning," "machine scor*" "machine rating," "automated rating," "automated scoring," and "Coh-Metrix." Thus, the final number of searches across databases was 162 (27 search combinations across 6 databases). An initial list of 193 items was compiled from the search results. These were selected by the researcher after briefly screening the items for relevance. An additional 26 items not found in the search were included as supplemental items. These were identified through references in the search result items and through independent research by the researchers. This total number of 219 articles, technical reports, and reviews was further reduced to 21 articles presenting or expanding upon empirical results. The reduction was based on the degree of focus on either AS of writing or speaking and this was enacted by the primary researcher upon a review of each article's abstract and findings. Ultimately, this resulted in 11 writing-focused articles and 10 speaking-focused articles. The researcher then completed annotated summaries for these. Moreover, five additional articles were identified as relevant literature reviews or surveys of the current state of AS in language assessment. These five items did not present new empirical findings. The result was that a final total of 26 primary items were selected to inform this review of AS in language assessment.

Key Themes in Automated Scoring of Language Proficiency in Speaking and Writing

Three key themes were identified based on the literature reviews and surveys of the current state of automated assessment. To identify themes, the researchers conducted a thematic analysis of the articles by first coding salient content in the articles and summaries, and then creating main themes and subthemes based on the code groups (Braun & Clarke, 2006). Tables 1 and 2 show how the three identified themes were distributed across the 11 writing-focused and 10 speaking-focused items. The first theme is AS design considerations, that is, a focus on the AS model design, operational practices, or purpose. The second theme is the role of humans and automation when considering both in relation to AS. In other words, this focus centred on how automation and humans combine before, during, and after the AS process to facilitate specialization, provide complementary support, influence scoring, and utilize AS results. The third theme is the accuracy of AS with different groups. Test-taking population differences related to first language (L1), gender, ethnicity, and country can occur between subgroups within the national level of a population or when comparing two different populations at an international level (Bridgeman et al., 2012b; Zhang et al., 2013). Thus, this includes discussion on AS accuracy, viability across population groups and specific populations, and the degree of generalizability. It should be noted that the third theme had less representation across the articles. However, as the examples that do occur relate to AS scalability, transferability, fairness, bias, and quality control as well as assurance, the researchers deemed that it was justifiable to include it as an important if somewhat comparably less salient theme. Further, the theme is addressed as an important one in surveys of the current state of AS (Rupp et al., 2020; Yan & Bridgeman, 2020).

Table 1

Writing Articles and Reports

Article title	Author(s)	Year	Theme 1	Theme 2	Theme 3
Automated subscores for TOEFL iBT® independent essays	Attali	2011	✓	✗	✓
Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country	Bridgeman et al.	2012b	✗	✓	✗
Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring	Attali et al.	2012	✗	✓	✗
Investigating the suitability of implementing the e-rater ® scoring engine in a large-scale English language testing program	Zhang et al.	2013	✓	✓	✓

Article title	Author(s)	Year	Theme 1	Theme 2	Theme 3
Monitoring of scoring using the e-rater® automated scoring system and human raters on a writing test	Wang & von Davier	2014	✓	✓	✗
Validity arguments for diagnostic assessment using automated writing evaluation	Chapelle et al.	2015	✓	✗	✗
The effect of using automated essay evaluation on ESL undergraduate students' writing skill	Aluthman	2016	✓	✓	✗
The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments	Wind et al.	2018	✓	✓	✗
Machine learning–driven language assessment	Settles et al.	2020	✓	✗	✗
More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms	Shin & Gierl	2021	✓	✗	✗
Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing	Latifi & Gierl	2021	✓	✗	✗

Note. Theme 1 = AS Design Considerations; Theme 2 = Role of Humans and Artificial Intelligence; Theme 3 = Accuracy of Automated Scoring with Different Groups

Table 2

Speaking Articles and Reports

Article title	Author(s)	Year	Theme 1	Theme 2	Theme 3
A comparison of two scoring methods for an automated speech scoring system	Xi et al.	2012	✓	✗	✗
TOEFL iBT speaking test scores as indicators of oral communicative language proficiency	Bridgeman et al.	2012b	✗	✓	✗
Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT™)	Zechner et al.	2015	✓	✓	✗

Article title	Author(s)	Year	Theme 1	Theme 2	Theme 3
Deep neural network acoustic models for spoken assessment applications	Cheng et al.	2015	✓	✗	✓
Comparative evaluation of automated scoring of syntactic competence of non-native speakers	Zechner et al.	2017	✗	✓	✓
Combining human and automated scores for the improved assessment of non-native speech	Yoon & Zechner	2017	✓	✓	✗
Monitoring the performance of human and automated scores for spoken responses	Wang et al.	2018	✗	✓	✗
Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models	Fu et al.	2020	✓	✗	✓
Detecting pronunciation errors in spoken English tests based on multifeature fusion algorithm	Wang	2021	✓	✓	✗
Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: A user perception study	Gu et al.	2021	✗	✓	✗

Note. Theme 1 = AS Design Considerations; Theme 2 = Role of Humans and Artificial Intelligence; Theme 3 = Accuracy of Automated Scoring with Different Groups

Automated Scoring Design Considerations

Numerous subthemes related to the automated scoring emerged. These subthemes are related to model design, model implementation considerations, and purposes.

Automated Scoring Model Design

Automated scoring models evaluate different language features (grammatical features, sophistication of vocabulary, usage errors, organization and discourse development) using algorithmic approximation of human rating. However, the processes used by human raters and AS differ. Where human raters match response features to rubrics, AS models extract proxies of rubric features from responses and use them in statistical models to yield predictive evaluations (Schneider & Boyer, 2020). This difference in evaluation processes is one of the main criticisms against AS. Namely, the concern is that AS may create validity issues by focusing on narrowed elements of language and that it may differ in a fundamental way from human rating and evaluation (Douglas, 2013; Schmidgall & Powers, 2017). A variety of AS statistical models can be used (multiple linear regression, lasso regression, nonnegative

least-squares regression, support vector machine, artificial neural network, deep learning, multimodal learning, best linear prediction, etc.); however, it is critical that model assumptions and abnormalities be examined in the contexts of use (Yan & Bridgeman, 2020).

In general, the use of Coh-Metrix approaches has a long track record in the research, whereas deep neural network approaches represent more state-of-the-art implementations. One advantage of the Coh-Metrix feature approach is that it requires smaller training sets than the deep-neural approaches. Moreover, Latifi and Gierl (2021) demonstrated a reasonable fit of Coh-Metrix models for a large-scale AS of essays. They showed that this traditional approach was adept at language capture and assessment across language features in essays. Thus, at present, there are advantages and disadvantages for each, which highlights the importance of selecting a scoring model design that suits the purpose and context of implementation. Both expert hand-crafted feature models and natural language processing feature-inducing AS models (i.e., using deep neural algorithms to learn the language features) are trained using human ratings (Hussein et al., 2019; Wind et al., 2018) or human-judgment informed materials (Settles et al., 2020).

How Automated Scoring Feedback is Perceived

Gu et al. (2021) looked at perceptions of using AS-based feedback for test preparation. Feedback types included domain subscores, task scores, and individual linguistic features. In general, teachers and test takers perceived the automated speaking feedback as being useful, though the teachers were more skeptical of its usefulness when compared with students. This may suggest that the teachers perceived limitations in the feedback that would otherwise benefit from teacher mediation, perhaps underscoring that automated feedback is not at a point where it can completely replace teachers. It may also suggest that the teachers were better able to evaluate the strengths and weaknesses of the feedback than were the test takers, at least in terms of their understanding of language and feedback usefulness. Nonetheless, automated feedback on speaking does seem to be perceived, by both teachers and learners, as being somewhat valuable for learning and teaching purposes.

Automated Scoring Implementation Focus

The focus of the scoring model can also vary significantly. For example, general scoring models can be deployed broadly and have the benefit of scalability, though they may rely on scoring surface-level language use and mechanics (Zhang et al., 2013). Scoring models can also be trained for more specific and focused use. As such, AS model designs need to be considered in relation to their design, performance with respect to the gold standard of qualified human raters, and specific contexts of use (Powers et al., 2015). Shin and Gierl (2021) explored the performance of a traditional automated essay scoring model (i.e., a model using support vector machines with Coh-Metrix features) and a deep neural model (convolutional neural networks) for scoring. The deep neural model performed better overall, though with some difficulties on specific types of constructed-response test items. This approach has been shown to produce assessments that correlate more highly with human raters, based on quadratic-weighted kappa comparisons, than the correlation between Coh-Metrix approaches and human raters.

Assessment Purpose and Item Complexity

Automated scoring—often of speaking and writing—can be used for a wide variety of purposes, including high-stakes purposes and formative feedback. This leads to varying levels of item complexity in the scoring process. For example, low complexity speaking items may require constrained or narrowed responses, or even reading from a script (Higgins et al., 2011). Conversely, open-constructed response items require AS scoring of unstructured, unrestricted, and spontaneous responses at different proficiency levels. Similarly, writing items may feature closed-item responses (cloze items, multiple choice, etc.) or open-constructed response items related to tasks, productions, or free responses (Hussein et al., 2019; Wang & von Davier, 2014), with the latter involving greater complexity. Examples of constructed response writing items include open-ended questions, short presentations, or short response interviews. Additionally, complex task-related items often involve multiple cognitive or psychological processes (reasoning, problem solving, arguing, etc.) and multifaceted output in terms of responses (Foltz et al., 2020). The increased complexity of constructed response items means there is a heightened need for quality assurance in terms of both AS validity and scoring capabilities (Chapelle et al., 2015; Wang & von Davier, 2014).

Stakes and Intended Use

In terms of purpose, there are examples supporting diagnostic, high- and low-stakes assessment, and formative purposes (Aluthman, 2016; Attali, 2011; Chapelle et al., 2015). As such, AS can be used in a wide variety of contexts, both high- and low-stakes (Wood, 2020). However, as the stakes increase, so does the reciprocal responsibility of assessment validity and fairness of AS (Rupp et al., 2020; Williamson et al., 2012); the design standards and infrastructure supporting implementation in high-stakes contexts must be heightened (Wood, 2020).

Automated scoring and feedback can also be used to support teaching and learning. Aluthman (2016) investigated the impact on student essay writing development over a long-term period of pedagogical support featuring automated feedback and teacher mediation. Students improved mainly in writing mechanics, with modest improvements in grammar, usage, and style. The scoring model identified and sorted elements of writing, and then supported the process of learning by highlighting these for students and giving feedback as well as providing iterative support to teaching by informing the teacher about student writing development and common problems. In general, the granularity of the AS model analysis influences the feedback capabilities (DiCerbo et al., 2020). As a wide spectrum of granularity is afforded across AS models, consideration of need with respect to granularity is important when feedback is the primary desired outcome. More detailed and granular feedback capabilities have greater potential formative use.

While large scale and high stakes AS use has enjoyed significant attention and application since its inception, formative and educational use of AS scoring and feedback has recently increased (Foltz et al., 2020). This increase is largely due to advances in technology and AS design (Foltz et al., 2020; Rupp et al., 2020). Though the precise balance of AS standards for specific contexts differs, it raises important issues of AS design transparency and public conversation (Wood, 2020).

Features Scored

The features of language scored among AS models differs significantly. In Wang (2021), a deep neural network model was created to model the feature values and to score speaking. A cutting-edge multidimensional feature extraction was performed on language recordings of test takers. Five categories of features were evaluated: pronunciation, fluency, vocabulary, grammar, and semantics. In Xi et al. (2012), two alternative automated speech scoring methods applied were compared: multiple regression and classification trees. The multiple regression model aligned more closely with the human scores and had greater construct relevance. In fact, the construct representation of the model was determined to meet the threshold needed for low-stakes test use of automated speech scoring. The speech scoring components included were automatic speech recognition, filters for flagging non-scorable responses, and linguistic measures of construct subdimensions, in addition to the two alternative scoring methods previously mentioned.

Improving Scorability

Further, Wang (2021) introduces text cleaning after speech recognition and deep learning-based noise reduction to improve speech recognition and scoring accuracy. These and other technological implementations suggest new possibilities in terms of open oral grading. Likewise, Yoon and Zechner (2017) used flagging with automated speech scoring, as well as automated speech feature recognition. That is, difficult-to-score items were flagged by an AS filtering system (using baseline and extended filters) and then scored by human raters. As such, this is an AS model design feature. This significantly improved system scoring correlation with human raters.

Scoring Different Task Types

Scoring of predictable and constrained spoken responses is well-established in the research and may have some advantages in terms of assessing speaking in and for specific constrained contexts and purposes (Litman et al., 2018). Tasks involving constrained assessment of speaking often include reading or production of an elicited response (Litman et al., 2018). Zechner et al. (2015) presented findings on successful AS of predictable and semi-predictable speech of speakers whose first language (L1) was not English. Cheng et al. (2015) investigated the effectiveness of a traditional Gaussian mixture model and a deep neural network hidden Markov model for acoustic modeling in educational applications of spoken assessment. The deep neural network significantly outperformed the traditional model. When comparing performance on open-ended and constrained tasks, the deep neural networks showed greater gains with the constrained tasks. Significant training data availability is needed for the deep neural network model training. Conversely, constrained spoken tasks require significantly less training data and have been proven reliable for constrained tasks.

Pronunciation

Fu et al. (2020) introduced an automatic proficiency evaluation system for the evaluation of pronunciation by applying a scoring system that included various non-L1 English speaker acoustic models and L1 English speaker models (Gaussian mixture model, hidden Markov model, and deep

neural network). They then introduced a novel machine score called the reference-free error rate to evaluate English proficiency without a specific reference anchor. Overall, the deep neural networks outperformed the traditional acoustic models.

Summary of Automated Scoring Design Considerations

In summary, the scoring model design of both speaking and writing has seen numerous advances. Traditional models have a proven fit for low-stakes assessment such as training, teaching, learning, and informal diagnostic assessment. Approaches leveraging cutting-edge technologies and scoring models (e.g., deep neural networks) appear to be nearing a tipping point of surpassing traditional models (e.g., Coh-Metrix) in both accuracy and viability for broad implementation. Nonetheless, the training data requirements, limitations on generalizability, and scaling still present challenges. Targeted and constrained use of traditional models still has significant utility. Hybrid artificial intelligence and human approaches may allow for controlled and strategic use of both established and cutting-edge AS models. Considering scoring model design, development, and implementation, as well as scoring model selection, a wide range of options are available. Still, stakeholders ought to choose wisely, based on their needs and capabilities (Williamson et al., 2012).

Human Involvement in Scoring and Artificial Intelligence

Scoring Roles

Attali et al. (2012) note the difficulties of machine scoring reliability when it comes to scoring complex and higher-order elements of writing. Moreover, their investigations into creating a hybrid approach with a division of focus—humans rating higher-order writing elements and the automated scoring model rating lower-order writing elements—highlighted the challenges of operationalizing hybrid approaches. As their investigation piloting a variety of hybrid-scoring model adjustments aimed at enhancing synergy between human raters and AS discovered, even slight changes can create unintended scoring imbalances. Human and automated essay rater scores using general scoring models are highly related on average and are similar across most subgroups. Additionally, operational policies and design can mitigate the impact of differences between human and machine raters reflected in reported scores (Bridgman et al., 2012b). Importantly, neither human raters nor automated writing scoring models perform without variation when applied broadly (Zhang et al., 2013). This suggests both limits on complexity in terms of generalizability and the important role of quality assurance and quality control in the scoring of writing. This last point is raised by Wang and von Davier (2014), who investigated methods for monitoring the scoring of written constructed responses by both human raters and AS models. They emphasized the need for monitoring the quality of scoring by both human raters and AS models across time, programs, and contexts. Wind et al. (2018) detailed the importance of considering and controlling how automated essay scoring models are trained using human ratings. This introduces yet another dynamic in AS—its human influence. Ultimately, it demonstrates that various problematic rater effects can be replicated by automated systems through the training process. As such,

AS model development and training must guard against undesirable rater effects during model design as well as quality assurance and quality control procedures.

In terms of rating speaking, human raters may yet enjoy some advantage in terms of their ability to evaluate more complex language elements (Bridgeman et al., 2012a). Automated speech scoring models (e.g., *SpeechRater*™) may be suitable for playing a role in reducing the burden, but, at least at this point, the role is complementary rather than replacing the need for any human rating. Zechner et al. (2017) presented research showing how when spoken responses were analyzed using an AS system, it was the part-of-speech element that correlated most closely with human ratings rather than the clause or phrase element. In a hybrid approach, human raters can also play a troubleshooting role to handle problematic items that the automated system identifies as being difficult for AS systems to score (Yoon & Zechner, 2017; Zechner et al., 2015). This can reduce the cost and demands of human scoring-related labour due to the scaling effect of automating the scoring process. Wang et al. (2018) describe processes using charts and evaluation statistics to monitor and evaluate the scoring of constructed responses by both human raters and AS models. The statistical monitoring proved useful for identifying outlier test items, human raters, and AS results. Though overall AS correlation with human raters was shown, variation with specific items can be problematic, thus highlighting the need for monitoring of AS, test items, human raters, and ongoing operationalization. Wang (2021) presents research on state-of-the-art automated spoken scoring which eliminates the need for experts to manually label keywords prior to scoring. This is an example of increased AS model independence and automatization. Of course, these kinds of shifts, increasing the automatization of scoring and reducing the transparency of the scoring process, must also be carefully balanced with informing stakeholders and skeptics about the AS systems and processes. Otherwise, resistance to AS can make implementation, use, adoption, and innovation challenging (Wood, 2020).

Summary of Human Involvement in Scoring and Artificial Intelligence

At present, the roles of human scorers, automated scoring models, test takers, learners, and teachers conform to a variety of patterns and dynamics based on scoring design, technology, and implementation in the assessment of both writing and speaking. These roles are changing rapidly, with automation's roles increasing at a pace commensurate with its new capabilities. Nonetheless, the idea that automated scoring occurs devoid of human involvement or influence is erroneous. At present, there appears to be a strong case for the strategic use of hybrid approaches in some cases.

Accuracy of Automated Scoring with Different Groups and Uses

Population Variation

Attali (2011) presents research detailing how an automated essay scoring system (i.e., *e-rater*™) that considers word choice, grammatical conventions within sentences, and fluency has been shown to be stable across major language groups. This is an important consideration for AS models that are intended to be used in large-scale testing and with diverse populations. Zhang et al. (2013) found that population factors can influence the scoring of some items scored by both humans and AS, though in

general, there is broad cross-population stability. In the case of AS, it seems that it can sometimes replicate human-rater variations in judgment. Using AS systems with diverse populations implies that the AS results of subgroups within these populations should show equal agreement with human raters for these subgroups (Yan & Bridgeman, 2020).

However, some construct-irrelevant linguistic and culture-related stylistic elements (e.g., shell language, discourse development or linearity, etc.) may have a minor, yet not insignificant, influence on scoring done by both AS and human raters (Bridgeman et al., 2012b; Yan & Bridgeman, 2020). Population variation in this section refers to differences of L1, gender, ethnicity, and country that may create minor effects—depending on AS design elements and test items—both when comparing different national populations and specific subgroups within countries (Bridgeman et al., 2012b; Zhang et al., 2013). Training automated systems with scoring data tuned to specific target test populations and languages can alleviate some of the problematic effects (Bridgeman et al., 2012b).

An example of human-rater variation by population provided by Zhang et al. (2013) is that the amount of shell language used in essays may be valued differently by different human raters based on their own writing feature expectations. Shell language is a general and non-specific sequence of words used in persuasive writing or speech to advance and frame an argument. The general nature of the word sequence allows it to be plugged into a wide range of persuasive contexts without direct construct relevance, whilst also increasing the overall wordiness of the language produced (Bejar et al., 2013). Depending on the scoring model and approach, wordiness and use of shell language may be valued differently. Moreover, different populations may be more or less likely to use shell language in essays, thus leading to variation in rating on some test items. These kinds of differences must be monitored and considered when using AS of writing broadly (Bridgeman et al., 2012b; Schneider & Boyer, 2020). Training data for AS come from specific populations or groups at specific points in time and the judgments these automatic scorings produce reflect this influence. As such, it is necessary to monitor AS for invariance, drift, and anomalies (Wang & von Davier, 2014). Fundamentally, AS of writing needs to fit the scoring demands of the context of use during model development or selection, training, and implementation. Without attention to these details, AS of writing is more vulnerable to variation effects.

Automated Scoring Fit to Populations and Contexts

Fu et al. (2020) introduce research detailing the challenges of assessing the pronunciation element in the speech of non-L1 English speakers. They used both L1 English and non-L1 English pronunciation models. Insufficient training of automatic speech scoring models on diverse pronunciation aspects of speech can limit the capabilities and generalizability of the models (Fu et al., 2020). Cheng et al. (2015) demonstrated a high level of scoring model performance using deep neural network-based scoring models and diverse speech production groups (e.g., children, adults, non-L1 English speakers, and L1 English speakers). Their investigations found that deep neural networks (DNN) outperformed a traditional AS model and that the word error recognition rate was significantly reduced, even when testing populations with significant variability. Test populations with significant variability in spoken language (children, non-native speakers of the language being tested, etc.) pose

challenges for AS and can result in word-recognition errors. A key point also noted in this investigation was the importance of the training data in DNN performance. DNN performance improves significantly with greater access to training data. Though the scoring of both constrained and open responses see improvement, open responses benefit most. Zechner et al. (2017) explore how to best evaluate syntactic competence on non-L1 English speakers using automation. They emphasize syntactic competence as being a key element of adept communication. Overall, these points suggest that using a scoring model within a specific domain or context may lend itself to more accurate scoring of constrained and predictable patterns of syntactic language use with a relatively predictable test-taking population (DiCerbo, 2020).

Summary of Accuracy of Automated Scoring with Different Groups and Uses

When compared to the scoring of writing, the scoring of speaking involves an additional challenge of correctly deciphering spoken utterances. This sometimes results in assessments constraining the language in tasks to expected and known patterns. Errors in transcribing the spoken language or limitations of AS judgment—based on training limitations, AS model design, or operational policies—pose real challenges for the automatic scoring of speech, and especially for unconstrained production from broad test-taking populations or specific populations that the model is not tuned to (D'Mello, 2020). Nevertheless, when mobilized with a targeted and constrained assessment focus and paired with careful AS model selection and policies, various elements of speech can be automatically scored in a reliable manner. The extent to which this can be fairly done with all language groups and users as well as with specific language groups and users is a critical question in the automatic assessment of speech at scale.

Automated Scoring Implementations for Assessment, Teaching, and Learning

Large-Scale and High-Stakes Testing

Given that some of the main forces behind the drive to use AS are the desire to reduce human labour costs, increase test security, and mobilize testing on a large scale, AS generalizability and validity—especially validity across populations—are critical considerations. This emphasizes the importance of AS model training. At present, state-of-the-art AS models require significant training data. Moreover, with high-stakes testing, the need for validity demands the highest level of AS quality assurance and quality control (Ricker-Pedley et al., 2020; Shaw et al., 2020). This high degree of quality assurance and control requires significant expertise in AS and language testing. As such, though cutting-edge AS appears to be on the cusp of being able to fulfill this large-scale and high-stakes testing promise, it still faces some critical hurdles. High-stakes AS testing may result in life-changing effects for test takers (program admissions, professional certification, immigration status, etc.), which underscores the importance of ensuring assessment language construct and feature validity as well as assessment security (Schmidgall & Powers, 2017). Examples of high-stakes AS of language include Test of English as a Foreign Language Computer-Based Test and Internet-Based Test, International

English Language Testing System, and Pearson Test of English Academic (Schmidgall & Powers, 2017).

The research presented in the previous sections suggests that though correlation with expert human raters is promising for technologically advanced AS, more testing and development is required to reach the threshold for cost-effective and high-quality testing that is appropriate for high stakes use on a large scale. Conversely, more traditional and established AS approaches have comparative limitations in terms of correlation with expert human raters, generalizability, and ability to handle language complexity. Less training data are required, but more human involvement is often needed.

The literature reviewed above emphasized the importance of communicating AS details and purposes to stakeholders in an accessible manner. As automated scoring is increasingly implemented in the context of high-stakes testing, this becomes more and more important (Wood, 2020). Without the buy-in of stakeholders, AS using both traditional and more cutting-edge methods may face resistance that could limit its development and potential.

Low-Stakes and Targeted Testing Within Specific Domains

In low-stakes contexts, AS has been used in practice tests, formative training, and for educational purposes in and out of classrooms (Foltz et al., 2020; Rupp et al., 2020; Shermis & Burstein, 2013). Research suggests that in contexts of targeted and known domains with semi-predictable language use, traditional AS, especially when paired with human raters in hybrid scoring approaches, have enough support to justify thoughtful implementation. These require smaller training sets than more cutting-edge AS models and it can be easier to design the assessment for specific populations. Moreover, specific language use can be targeted within the domain. This has benefits for focusing the AS model, accurately reflecting valid constructs of language use, and predicting language output. In this context, AS use for low-stakes assessment and training can be sufficiently accurate, cost-effective, and scalable for broad specific use within domains. With reduced stakes, the validity threshold can also be lowered somewhat since the purpose of the AS has less potential negative impact on user outcomes within the context of AS use.

Formative Feedback for Teaching and Learning

This research literature shows that both writing and speaking feedback stemming from traditional AS can be used for pedagogical purposes. The feedback capabilities of an AS system depend not only on building feedback into the design, but also on the level of fine-grained analysis and detail in the design (DiCerbo et al., 2020). This is significant as it is an important AS design consideration if feedback is to be effectively transferred to support teaching and learning.

Teacher-mediated use could involve long-term monitoring of student development and needs, as well as integrated pedagogical use. While there may be limitations in terms of the language focus and complexity that the feedback can target, it nonetheless has a well-established history of practical and valid use.

Teacher-unmediated use of AS feedback also has demonstrated usefulness as a learning support (Burstein et al., 2021; Fu et al., 2020). Though it is not a "magic bullet" capable of replacing teacher mediation, it does have an established footing as a language learning support element that is operational at a very broad scale (Loewen et al., 2019). Automated scoring feedback for independent learning offers some interesting possibilities related to both traditional AS approaches and more state-of-the-art approaches. Given the pandemic-related assessment, teaching, and learning challenges, ways to mobilize AS feedback and remote learning access ought to be actively pursued.

Future Research on Automated Scoring of Speaking and Writing

Formative Purposes and Narrow Context of Language Use and Proficiency

As was briefly discussed in the previous subsection, AS can play a formative role in teaching and learning. More research on the long-term impact of AS implementation for formative language learning purposes, especially with state-of-the-art AS models, is needed. The case for AS usefulness in this context is well-established, but further exploration of teaching and learning would help to refine the implementation. This should include use in both synchronous and asynchronous teaching contexts, remote teacher-mediated contexts, and remote teacher-unmediated contexts. Providing iterative feedback and support based on ongoing AS model judgments may offer significant social benefits in terms of both teaching and learning. Additionally, cutting-edge AS use opens doors to the leveraging of metadata for teaching and learning and feedback on more complex language elements. Of course, this also heightens the fiduciary duty of AS developers and requires proactive outreach related to informing stakeholders and operationalizing AS approaches with appropriate transparency.

Domain-Specific and Population-Specific Automated Scoring

More research on tailored AS models for specific domains and populations is also needed. Using both traditional and cutting-edge AS models, targeting AS model training and development for specific populations and contexts can increase validity and fairness. That is, fairness is increased by accounting for specific predicted populations of users and possible outlier users (e.g., those with different language, culture, or digital literacy). Validity is increased by selecting language constructs that are appropriate for specific domains. In both cases, AS models may benefit from being trained on data sets that closely resemble the future participants in terms of language production. This has an additional potential benefit of including practical domain information that may be relevant to academic or professional competencies. For example, an AS approach for nurses might feature the inclusion of tasks featuring relevant writing conventions or verbally relaying general health-related information. This type of tailored AS development also happens to be a good candidate for formative AS mobilization. As mentioned previously, domain-specific AS applications for formative learning may allow for the harvesting of informative metadata that could further deepen our understanding of the language learning process and the role of AS-based feedback.

Next Steps

Automated scoring technological capabilities continue to grow, as does our understanding of how to implement them effectively in language assessment and testing. Automated scoring of speaking and writing, the productive language skills, have shown significant development. These language skills are critical for social and professional engagement in today's increasingly globalized world. With this in mind, it is important to consider the future role AS will play in Canadian society. Providing increased access to cost-effective, efficient, and reliable language assessment, teaching, and learning is of paramount importance. Remote AS assessment, teaching, and learning options not only lead to increased access, both generally and considering specific pandemic-related challenges, but also reduce costs and service bottlenecks. Reducing assessment bottlenecks whilst providing accessible teaching and learning supports can facilitate increased accreditation of internationally trained professionals and increase their participation in Canadian society. Language proficiency appropriate for the Canadian context is critical for professional success and related language proficiency limitations have been identified as a significant barrier for internationally trained professionals in Canada (Kaushik & Drolet, 2018). Barriers to professional accreditation have negative effects on both the individuals, who are at risk of being marginalized, and Canadian society, which is both in dire need of skilled labour and increased socioeconomic inclusion of marginalized people (Kaushik & Drolet, 2018). Automated scoring is poised to play an important role in overcoming these challenges. It is starting to hit its stride and will continue to gain speed based on new technologies, new applications, and new research.

References

Aluthman, E. S. (2016). The effect of using automated essay evaluation on ESL undergraduate students' writing skill. *International Journal of English Linguistics*, *6*(5), 54-67. https://doi.org/10.5539/ijel.v6n5p54

Attali, Y. (2011). Automated subscores for TOEFL iBT® independent essays. (ED525308). *ETS Research Report Series*, *2011*(2), i-16. https://doi.org/10.1002/j.2333-8504.2011.tb02275.x

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125-141. https://doi.org/10.1177/0265532212452396

Bejar, I. I., VanWinkle, W., Madnani, N., Lewis, W., & Steier, M. (2013). Length of textual response as a construct-irrelevant response strategy: The case of shell language. *ETS Research Report Series*, *2013*(1), i-39. https://doi.org/10.1002/j.2333-8504.2013.tb02314.x

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012a). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, *29*(1), 91-108. https://doi.org/10.1177/0265532211411078

Bridgeman, B., Trapani, C., & Attali, Y. (2012b). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2021). A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English test. http://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem.pdf

Cahill, A., & Evanini, K. (2020). Natural language processing for speaking and writing. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 69-92). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

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

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication*, *73*, 14-27. https://doi.org/10.1016/j.specom.2015.07.006

D'Mello, S. (2020). Multimodal analytics for automated assessment. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 93-111). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

d'Orville, H. (2020). COVID-19 causes unprecedented educational disruption: Is there a road towards a new normal? *Prospects*, *49*, 11-15. https://doi.org/10.1007/s11125-020-09475-0

DiCerbo, K., Lai, E., & Ventura, M. (2020). Assessment design with automated scoring in mind. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 29-47). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Douglas, D. (2013). Technology and language testing. In C. A. Chapelle (Eds.), *The encyclopedia of applied linguistics* (pp. 1-7). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1182

Foltz, P. W., Yan, D., & Rupp, A. A. (2020). The past, present, and future of automated scoring for complex tasks. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 1-11). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, *116*, 86-97. https://doi.org/10.1016/j.specom.2019.12.002

Golkova, D., & Hubackova, S. (2014). Productive skills in second language learning. *Procedia-Social and Behavioral Sciences*, *143*, 477-481. https://doi.org/10.1016/j.sbspro.2014.07.520

Gu, L., Davis, L., Tao, J., & Zechner, K. (2021). Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: A user perception study. *Assessment in Education: Principles, Policy & Practice*, *28*(1), 58-76. https://doi.org/10.1080/0969594X.2020.1735995

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*(2), 282-306. https://doi.org/10.1016/j.csl.2010.06.001

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Kaushik, V., & Drolet, J. (2018). Settlement and integration needs of skilled immigrants in Canada. *Social Sciences*, *7*(5), 76. https://doi.org/10.3390/socsci7050076

Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, *38*(1), 62-85. https://doi.org/10.1177/0265532220929918

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, *15*(3), 294-309. https://doi.org/10.1080/15434303.2018.1472265

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293-311. https://doi.org/10.1017/S0958344019000065

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy*, *4*(4), 351-370. https://doi.org/10.1007/s10993-005-2886-0

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education*, *28*(2), 130-142. https://doi.org/10.1080/08957347.2014.1002920

Ricker-Pedley, K., Hines, S., & Connolley, C. (2020). Operational human scoring at scale. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 171-193). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Rupp, A., Foltz, P., & Yan, D. (2020). Theory into practice: Reflections on the handbook. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 475-487). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*(4), 302. https://doi.org/10.1037/0003-066X.56.4.302

Schmidgall, J. E., & Powers, D. E. (2017). Technology and high-stakes language testing. In C. A. Chapelle, & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 317-331). Wiley Blackwell. https://doi.org/10.1002/9781118914069.ch21

Schneider, C., & Boyer, M. (2020). Design and implementation for automated scoring systems. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 217-239). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247-263. https://doi.org/10.1162/tacl_a_00310

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge Academic.

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247-272. https://doi.org/10.1177/0265532220937830

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication*, *13*(2), 225-236. https://doi.org/10.1080/14708477.2013.770868

Voogt, J., & Knezek, G. (2021). Teaching and learning with technology during the COVID-19 pandemic: Highlighting the need for micro-meso-macro alignments. *Canadian Journal of Learning and Technology*, *47*(4). https://doi.org/10.21432/cjlt28150

Wang, Y. (2021). Detecting pronunciation errors in spoken English tests based on multifeature fusion algorithm. *Complexity*, *2021*, 1-11. https://doi.org/10.1155/2021/6623885

Wang, Z., & von Davier, A. A. (2014). Monitoring of scoring using the e- rater® automated scoring system and human raters on a writing test. *ETS Research Report Series*, *2014*(1), 1-21. https://doi.org/10.1002/ets2.12005

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101-120. https://doi.org/10.1177/0265532216679451

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. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wind, S. A., Wolfe, E. W., Engelhard Jr, G., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, *18*(1), 27-49. https://doi.org/10.1080/15305058.2017.1361426

Wood, S. (2020). Public perception and communication around automated essay scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 133-150). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371-394. https://doi.org/10.1177/0265532211425673

Yan, D., & Bridgeman, B. (2020). Validation of automated scoring systems. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 297-318). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781351264808

Yoon, S. Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, *93*, 43-52. https://doi.org/10.1016/j.specom.2017.08.001

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S. Y. (2015). Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT™). *ETS Research Report Series*, *2015*(2), 1-17. https://doi.org/10.1002/ets2.12080

Zechner, K., Yoon, S. Y., Bhat, S., & Leong, C. W. (2017). Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Computers in Human Behavior*, *76*, 672-682. https://doi.org/10.1016/j.chb.2017.01.060

Zhang, M., Breyer, F. J., & Lorenz, F. (2013). Investigating the suitability of implementing the E-Rater® scoring engine in a large-scale English language testing program. *ETS Research Report Series*, *2013*(2), i-60. https://doi.org/10.1002/j.2333-8504.2013.tb02343.x

Authors

Daniel Marc Jones is a PhD student in the Faculty of Education, Queen's University. His research focuses on the use of games to teach language and literacies with the pedagogy of multiliteracies. That is, games are framed as rich cultural artifacts that can support active learning and participatory opportunities. Email: 20dmj@queensu.ca

Liying Cheng is Professor and Director of Assessment and Evaluation Group (AEG) at the Faculty of Education, Queen's University. Her seminal research on washback illustrates the global impact of large-scale testing on instruction, the relationships between assessment and instruction, and the academic and professional acculturation of international and new immigrant students, workers, and professionals in Canada. Email: liying.cheng@queensu.ca https://orcid.org/0000-0002-4458-5085

M. Gregory Tweedie is Associate Professor in Language & Literacy at the Werklund School of Education, University of Calgary, Alberta. His teaching and research, in the field of applied linguistics, focuses on what happens when people from different first language backgrounds use English as a communicative vehicle in international professional contexts. Email: gregory.tweedie@ucalgary.edu.qa https://orcid.org/0000-0003-0497-4577