

Score Comparability Limits with Common AES Validation Measures

Tim Moses
Buros Center for Testing

Seongeun Kim
National Commission on
Certification of Physician Assistants

Nick Trout
Michigan State University

Abstract

This study considers measures commonly used to validate AES scoring and their limitations for indicating comparability with human rater scoring. In data simulated to reflect validation results achieved by AES national competition winners, scoring standard differences can occur across AES and human rater scoring (especially for 4-point scales vs. 2- and 3-point scales). Equipercentile methods are described and recommended for resolving the scoring differences.

1 Introduction

Automated Essay Scoring (AES) is often implemented within hybrid scoring systems, where some test takers' written responses are scored with AES and others are scored by humans (Kwako, Lottridge, & Ormerod, 2026; Palermo & Wibowo, 2024). The human scoring in hybrid systems fulfills at least two roles, to score written responses determined to be scored with difficulty or "low confidence" by AES methods, and to provide validation checks for the AES scores. To the extent that human scoring is considered the "gold standard" for the validation and interpretation of automated scoring (Palermo & Wibowo, pp. 23-24), human-machine agreement and associated measures are critical for AES validation.

What measures of human-machine agreement are most important for the validation of automated scoring? The most frequently used and reported indices include quadratic weighted kappas (QWKs, Cohen, 1968), average squared differences between scores rated by human and machine (Ramesh & Sanampudi, 2022), exact agreement percentages, and standardized mean differences (SMDs; Williamson, Xi, & Breyer, 2012).

Typically, these measures are evaluated with respect to pre-defined thresholds, such as 0.70 (70%) or higher for QWKs and/or score agreements, and 0.10 or 0.15 for standardized mean differences. Although these measures address several aspects of human-machine scoring agreement, the QWK is increasingly criticized for issues like estimation sensitivity to rating scales, issues with the marginal score distributions, and inconsistencies with alternative agreement measures (Doewes, Kurdhi, & Saxena, 2023; Lewis & Casabianca, 2026). One aspect of the measures, their criticisms and proposed alternatives do not directly address is scoring comparability, or the extent to which scoring standards are comparable across the two sets of scores (i.e., human and AES).

Potential scoring standard differences could be present and not well detected in QWKs of any sets of scores (AES vs. human, human vs. human, etc.) However, these limitations are especially pronounced in AES vs. human scoring, where AES scores can exhibit reduced spreads (fewer extreme scores) and more systematic differences as compared to human vs. human disagreements that might be more random. To the extent that AES scoring is implemented in hybrid systems with no validation and adjudication procedures typically used with human scoring, any comparability issues in AES vs. human scoring that are undetected in AES validation measures like QWKs are especially consequential for entire testing populations.

The purpose of this study is to examine the limitations of common AES validation measures for ensuring comparability across AES and human scoring. For a range of rating scales (2-, 3-, and 4-point scales), we consider how different scoring standards can be for AES models that have won national competitions (Shermis, 2024). Reported

score statistics, QWKs and % exact agreements for these competition winners are reconsidered with respect to how different or incomparable the scoring standards can be. In the field of linking and equating, the primary measure for establishing comparability is when two sets of scores have equal cumulative distributions (Moses, 2025). In this study, we analyze how different the cumulative distributions of the simulated can be to reflect the human and AES score characteristics for national winners of AES scoring model competitions. We also describe and demonstrate equipercentile conversions for resolving the scoring differences shown in the simulated data as potential impacts in seemingly validated AES scoring systems.

2 Methods and Samples

The questions of interest are examined using simulated discrete distributions. We obtained the human and AES score means, standard deviations, QWKs and % exact agreement from the winners of a recent Automated Student Assessment Prize (ASAP) competition (Shermis, 2024, Tables 7.3 & 7.4) for 20 2-, 3- and 4-point essays. We simulated discrete distributions to match the scales and statistical characteristics of these essays, and also to maximize a measure of (in)comparability in the human and AES scores. The measure we considered for score incomparability is defined as an average absolute difference in cumulative proportions of human and AES scores,

$$\text{Avg. Abs. CProportionDifference} = \sum_j p_{j, \text{Human}} |p_{j \leq j, \text{Human}} - p_{j \leq j, \text{AES}}| \quad (1)$$

where the p 's indicate proportions of test takers with scores in the complete set of J scores less than or equal to score j . Equation 1 indicates the extent to which cumulative proportions of test takers' AES and human scores at each possible score differ in absolute value, averaged across the distribution of human scores. For descriptive purposes, the maximum absolute difference is also considered,

$$\text{Max. Abs. CPercentageDifference} = 100\{\text{Max}[p_{j \leq j, \text{Human}} - p_{j \leq j, \text{AES}}]\} \quad (2)$$

To the extent that scoring standards differ (e.g., higher scores are more difficult to achieve with AES vs. humans and exhibit larger cumulative proportions), values for Equations 1 and 2 will be larger and more indicative of different or

incomparable scoring across AES and human systems. By generating bivariate distributions that exhibited the agreements achieved in national AES scoring competitions and also maximize Equation 1, we could consider how incomparable scoring standards could get when AES scoring that meets validation thresholds and wins national competitions.

To the extent that AES scores and human rater scores are systematically different and reflect differences in leniency or stringency, the raw AES and human scores should have at least some score points with different cumulative percentages. These differences in cumulative percentages indicate that the same score point may represent a different relative position in the AES and human score distributions. As mentioned, two sets of scores are considered comparable when they have equal cumulative distributions in linking and equating (Moses, 2025). Therefore, to address differences in the cumulative distributions of AES and human rater scores, equipercentile conversions were produced. Specifically, equipercentile conversions of AES scores to the human score scales are computed by identifying scores on the human scale with the same cumulative percentages as the corresponding AES scores (approximately "equal" percentiles).

3 Results

Table 1 presents overall results, including the human and AES score statistics and agreement measures from Shermis (2024), and the measures of scoring incomparability (Equations 1 and 2). Results show that for national competition winners with QWKs almost always above established thresholds (>0.70) and with relatively small differences in mean scores and SMDs, consequential differences in scoring standards can occur. The potential for scoring standard differences is up to 3% for essays with 2-point scales, 6% for essays with 3-point scales, and 7% for essays with 4-point scales. We consider both proportions and cumulative proportions of scores of the score distributions in detail for all essay items, depicted in Figures 1-3.

Table 1. Human and AES score statistics and agreement measures for national AES competition winners from Shermis (2024), and Avg.Abs.CProportionDifferences and Max.Abs.CPercentageDifferences.

Item	From Shermis (2024)							Current Study	
	Score Range	Mean (Human)	SD (Human)	Mean (AES)	SD (AES)	%Exact Agreement	QWK	Avg. Abs. CProporitonDifference	Max. Abs. CPercentageDifference
DR04 1715RE1T10 05	1-2	1.23	0.42	1.24	0.43	0.95	0.88	0.01	1%
DR04 1715RE4T05G04 07	1-2	1.45	0.50	1.46	0.50	0.94	0.87	0.01	1%
DR08 1715RE2T13 05	1-2	1.66	0.48	1.69	0.46	0.91	0.79	0.01	3%
DR08 1715RE4T05G08 07	1-2	1.80	0.40	1.81	0.39	0.94	0.80	0.00	2%
DR04 1715RE4T05G04 03	1-3	1.55	0.73	1.57	0.74	0.89	0.90	0.01	2%
DR04 1715RE4T05G04 09	1-3	1.55	0.67	1.55	0.67	0.78	0.75	0.01	1%
DR04 1715RE4T08G04 03	1-3	1.96	0.71	1.95	0.70	0.91	0.89	0.01	2%
DR04 1715RE4T08G04 06	1-3	1.56	0.80	1.57	0.80	0.96	0.97	0.01	1%
DR04 1715RE4T08G04 09	1-3	1.46	0.57	1.48	0.56	0.87	0.80	0.02	3%
DR08 1715RE2T13 08	1-3	1.99	0.73	2.02	0.73	0.75	0.69	0.01	2%
DR08 1715RE4T05G08 03	1-3	2.22	0.88	2.22	0.89	0.90	0.92	0.01	1%
DR08 1715RE4T05G08 09	1-3	2.20	0.75	2.15	0.72	0.81	0.81	0.02	6%
DR08 1715RE4T08G08 03	1-3	2.46	0.69	2.44	0.68	0.92	0.89	0.01	2%
DR08 1715RE4T08G08 06	1-3	2.42	0.85	2.41	0.85	0.96	0.97	0.00	1%
DR08 1715RE4T08G08 09	1-3	1.94	0.66	1.97	0.63	0.81	0.77	0.01	4%
DR04 1715RE4T05G04 06	1-4	1.90	1.05	1.90	1.06	0.82	0.90	0.03	7%
DR04 1715RE4T08G04 07	1-4	2.02	0.99	2.01	0.99	0.94	0.97	0.02	4%
DR08 1715RE2T13 07	1-4	1.65	0.84	1.68	0.91	0.83	0.89	0.03	5%
DR08 1715RE4T05G08 06	1-4	2.86	1.11	2.90	1.10	0.77	0.87	0.01	2%
DR08 1715RE4T08G08 07	1-4	3.02	0.95	3.03	0.95	0.93	0.96	0.01	3%

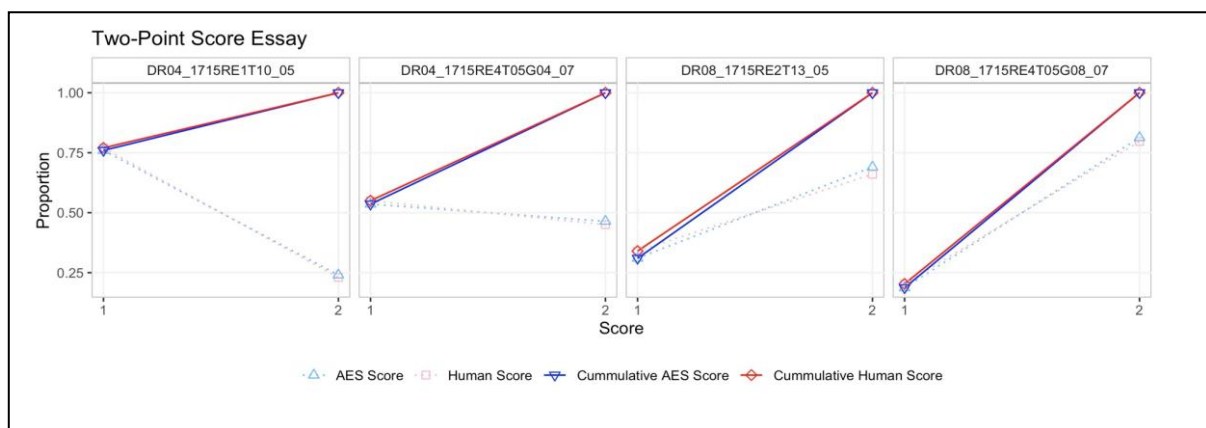


Figure 1. AES and Human Score Distributions (Essays with 2-point scales).

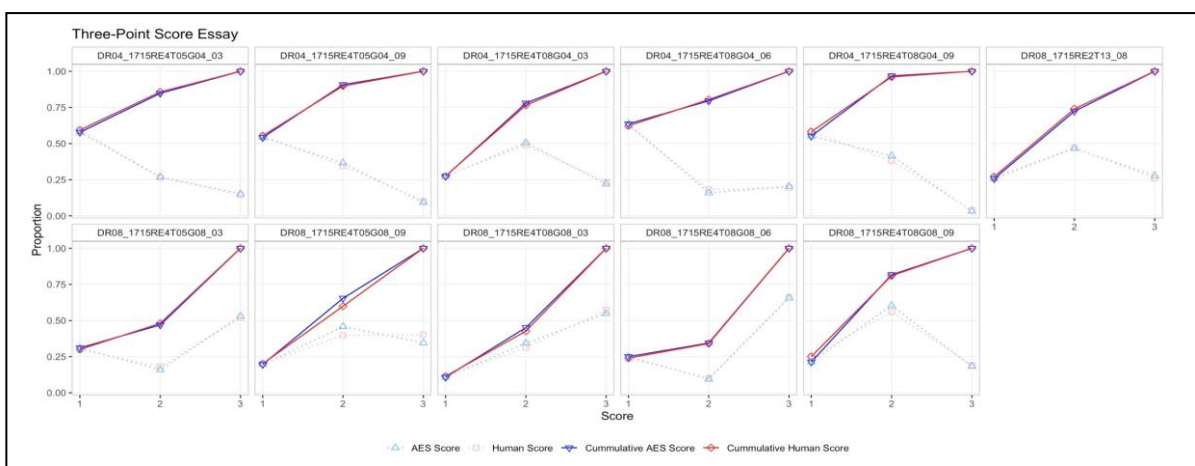


Figure 2. AES and Human Score Distributions (Essays with 3-point scales).

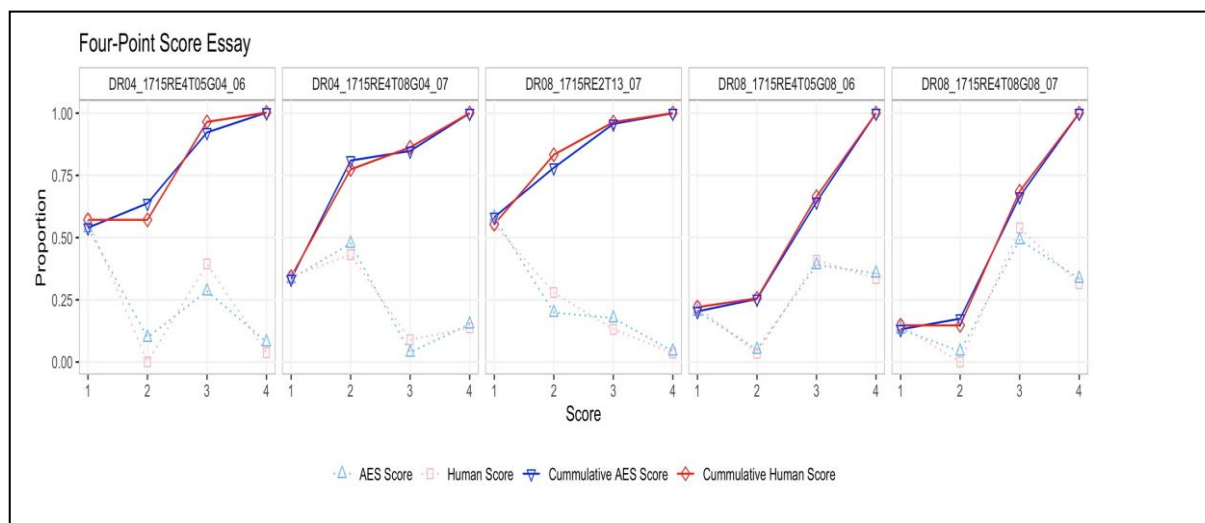


Figure 3. AES and Human Score Distributions (Essays with 4-point scales).

Overall, the proportions and cumulative proportions of AES and human scores in Figures 1-3 show two patterns. First, differences between AES and human score distributions were more apparent as the number of score points increased. Proportions and cumulative proportions were more similar for 2- and 3-point essays, and exhibited greater differences for 4-point essays. Second, human ratings tended to be more stringent than AES ratings, although the pattern varied across score categories. For the 2-point essays, human raters assigned a greater proportion of scores to 1 than AES, whereas AES assigned a greater proportion to score 2. For the 3-point essays, AES assigned a greater proportion of scores to higher score categories than human raters. For the 4-point essays, the differences between human and AES ratings varied across score categories. Between the two extreme score categories (1 and 4), AES assigned a greater proportion of scores to score 4, whereas between the two middle score categories, AES assigned a greater proportion of scores to score 2.

As previously mentioned, two sets of scores are considered comparable when their cumulative distributions are equal (Moses, 2025). Therefore, AES scores were linked to human score scales, reflecting a treatment of human scores as the gold standards. These linkings were computed as unsmoothed equipercentile conversions. The equipercentile results are summarized in Figures 4-6, which show the differences in the AES-to-Human equipercentile converted scores and the Human scores across the essay score scales. Consistent with the differences observed in the AES and human score distributions (Figures 1-3), the average absolute difference between AES-to-Human converted scores and Human scores increased as the number of score points increased: from 0.022 for the 2-point essays, to 0.029 and 0.135 for the 3- and 4-point essays. As expected, when AES scores were converted to the scales of the human raters, the converted AES scores tended to be lower than the raw and unconverted AES scores. The differences were relatively small for the 2- and 3-point essays but were more apparent for the 4-point essays. In particular, differences greater than 0.5 score points were observed for the first (DR04_1715RE4T05G04_06) and last (DR08_1715RE4T08G08_07) 4-point essays. For these items, the proportions of responses assigned a score of 2 by human raters were close to 0%,

which likely contributed to the relatively large converted score differences.

4 Conclusions

This study contributes to discussions about limitations of commonly used validation measures for AES (Doewes, Kurdhi, & Saxena, 2023; Lewis & Casabianca, 2026). For AES systems that meet thresholds for agreement and even win national score model competitions, meaningful differences in AES vs. human scoring standards can occur. These score comparability issues have implications for AES uses in the field, including hybrid scoring systems for state testing where AES scoring can be shown to meet established validation thresholds with respect to human scoring but still exhibit comparability issues like few or no highest possible scores, different score variances, or other scoring discrepancies.

These findings highlight the need to consider the magnitude of scoring (in)comparability directly, such as through distributions and equipercentile conversions, even when agreements exceed acceptable thresholds for AES validation measures (e.g., QWKs, % exact agreements, and SMDs). As illustrated in Figures 1-6, even when QWKs were high enough to win AES competitions, differences could be observed in the distributions of AES and human rating scores (Figures 1-3), and in equipercentile score conversions (Figures 4-6). These discrepancies tended to be larger as the number of score points increased, especially for essays with 4-point scales. The larger differences could be practically meaningful and affect score interpretations. Therefore, it is important to not rely solely on QWKs but also to consider the magnitude of scoring differences at each score point. A final implication of our results is that unrounded AES and human score scales may be useful to consider for hybrid scoring systems. Scoring differences of about 0.2 score points were found in some of our equipercentile results. Although these differences may not be observed in scores reported after rounding, they are large enough that discrepant rounded score distributions could occur. Unrounded equipercentile converted scores may be especially useful for the reporting of essay scores in hybrid scoring systems, and for addressing AES vs. human scoring differences not indicated in frequently used AES validation measures.

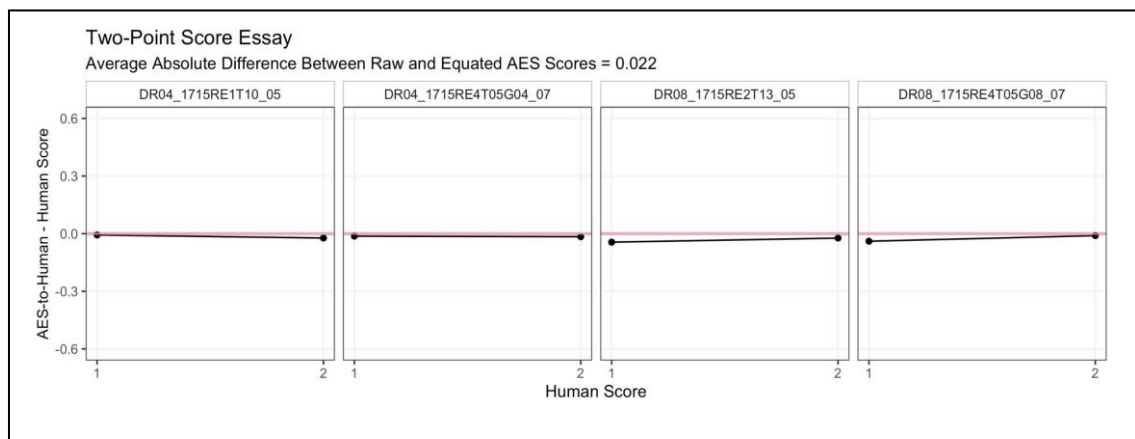


Figure 4. Differences in AES-to-Human Equipercentile and Human Scores (Essays with 2-point scales).

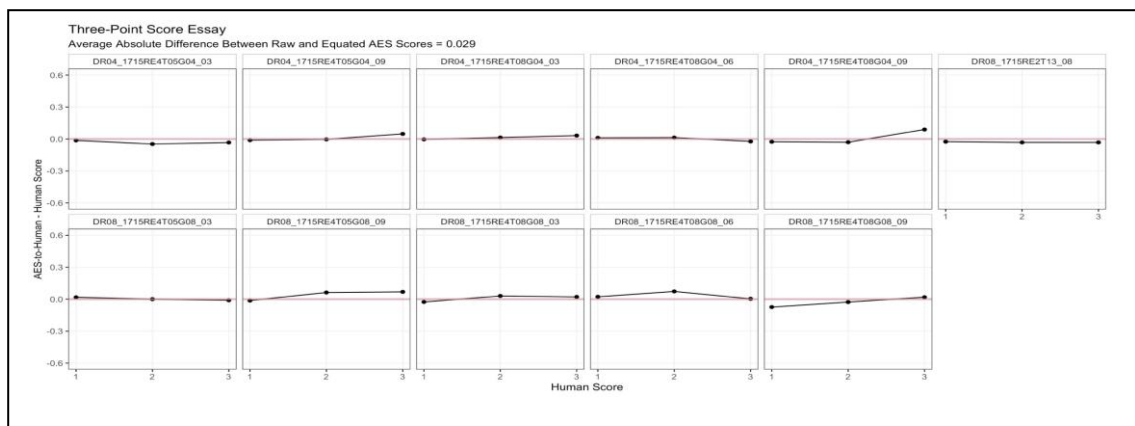


Figure 5. Differences in AES-to-Human Equipercentile and Human Scores (Essays with 3-point scales).

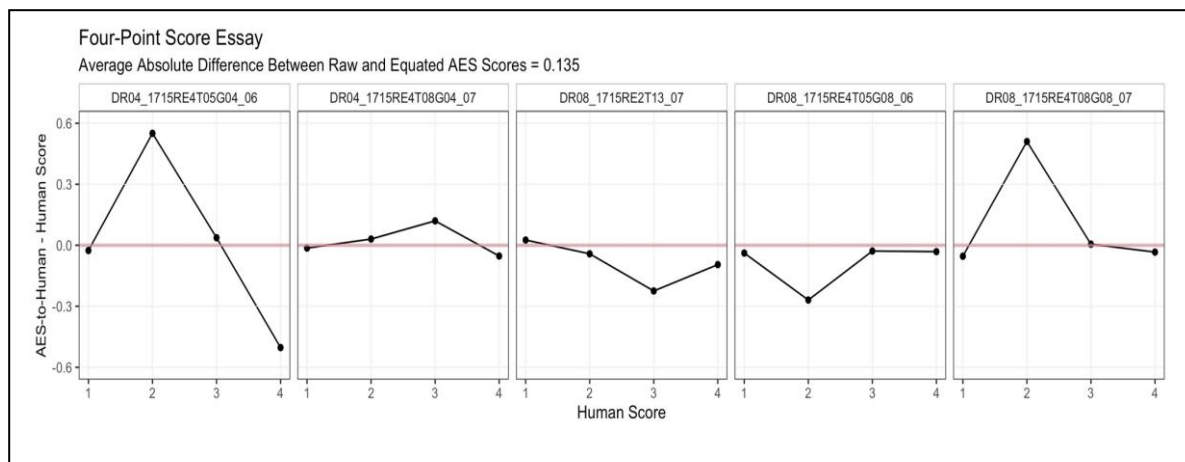


Figure 6. Differences in AES-to-Human Equipercentile and Human Scores (Essays with 4-point scales).

References

- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213-220. <https://doi.org/10.1037/h0026256>
- Chung, K. W. K. & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays* (ERIC reproduction service no ED 418 101).
- Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). Evaluating quadrative weighted kappa as the standard performance metric for automated essay scoring. In M. Feng, T. Kaser, & P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 101-113). International Educational Data Mining Society (IEDMS). <https://doi.org/10.5281/zenodo.8115784>
- Huang, Y., Palermo, C., & Wilson, J. (2026). Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing*, 69, 101047. <https://doi.org/10.1016/j.asw.2026.101047>
- Kwako, A., Lottridge, S., & Ormerod, C. (2026). Using confidence modeling to optimize overall score quality in hybrid scoring systems. *Educational Measurement: Issues and Practice*, 45(2), <https://doi.org/10.1111/emip.70019>
- Lewis, J., & Casabianca, J. M. (2026). Limitations of QWK in evaluating automated and human scoring systems. *Educational Measurement: Issues and Practice*, 45(1), <https://doi.org/10.1111/emip.70017>
- Moses, T. (2025). Scaling, equating, and linking. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 735-820). National Council on Measurement in Education. <https://doi.org/10.1093/oso/9780197654965.002.0003>
- Palermo, C., & Wibowo, A. (2024). Automated essay evaluation at scale: Hybrid automated scoring/hand scoring in the summative assessment program. In Shermis, M. D., & Wilson, J. (Eds.). *The Routledge International Handbook of Automated Essay Evaluation* (pp. 23-39). Routledge. <https://doi.org/10.4324/9781003397618-3>
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial intelligence review*, 55(3), 2495-2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring, Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25-39.
- Shermis, M. D. (2024). Automated scoring for NAEP short-form constructed responses in reading. In Shermis, M. D., & Wilson, J. (Eds.). *The Routledge International Handbook of Automated Essay Evaluation* (pp. 117-140). Routledge. <https://doi.org/10.4324/9781003397618-9>
- 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>