Comparative Study of Double Scoring Design for Measuring Mathematical Quality of Instruction

Jonathan K. Foster[1], James Drimalla[2], Nursultan Japashov[1]
[1]University at Albany
[2]Gordon College
jkfoster@albany.edu, james.drimalla@gordon.edu, njapashov@albany.edu

Abstract

The integration of automated scoring and addressing whether it might meet the extensive need for double scoring in classroom observation systems is the focus of this study. We outline an accessible approach for determining the interchangeability of automated systems within comparative scoring design studies.

1 Introduction

Classroom observation instruments may be deployed in different *classroom observation systems*, i.e., the collection of elements that work together to produce instructional quality ratings such as the observation instrument, raters, and scoring design (Hill et al., 2012). Classroom observation systems operating within education research or large-scale operational use have different goals and constraints than those operating for practical judgements on instructional quality (Liu et al., 2019). For instance, some classroom observation systems embedded in educational research may need calibration and monitor ratings, double scoring of observations, and complete multiple observations of teachers whereas classroom observation systems embedded in a large school district may not match all of these elements. Recent research highlights the need for extensive double scoring to determine whether raters are scoring accurately and consistently (White and Ronfeldt, 2024).

One potential approach to address the extensive need for double scoring is to pair human raters with an automated scoring system (Rotou and Rupp, 2020; Rupp, 2018). In recent years, a growing number of machine learning techniques have been used to identify features of instructional quality in classrooms from videos and audio recordings, or classroom transcripts. In one such study, researchers explored the zero-shot performance of ChatGPT (`gpt-3.5-turbo`) in scoring transcript segments from 4th- and 5th-grade mathematics instruction by applying the Mathematical Quality of Instruction (MQI) tool, a classroom observation instrument (for more information about MQI, see Hill et al., 2008). Results indicated the Spearman correlation between human and machine ratings for dimensions of MQI were low (Wang and Demszky, 2023). In another study, researchers applied a multimodal model and ChatGPT (`gpt-3.5-turbo-1106` and `gpt-4-1106-preview`) to video, audio, and transcripts to score encouragement and warmth in classrooms, a key component of the Global Teaching Insights (GTI) study's observation protocol (Hou et al., 2024). They found pairing the multimodal model with ChatGPT-4 yielded a moderate Pearson correlation ($r = 0.513$). Studies such as these illustrate the opportunities for automated scoring systems in classroom observation.

Current research investigating these automated scoring systems for classroom observation have primarily compared the performance of the automated system to that of human ratings. In terms of automated scoring systems, this focus is one of several components in an argument-validity framework (Rotou and Rupp, 2020; Williamson et al., 2012). These systems depend on human scoring for development. Yet, some scholars critique the lack of theoretical attention to measurement and reporting of inter-rater reliability for classroom observations and question whether classroom

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 163–171

observation systems that rely only on human raters can even consistently and accurately measure instructional quality (Liu et al., 2019; White and Ronfeldt, 2024; Wilhelm et al., 2018). Rather than shy away from these complexities with classroom observation systems or call into question the conclusions of some of the recent research on automated scoring systems for classroom observation, we propose an approach to guide others in this area in reporting their results.

The purpose of this paper is to examine an approach for illustrating the implications for double scoring in classroom observation systems when one of the raters is an automated scoring system and the other is a human, especially in the context of smaller datasets for initial system development. We make use of a dataset from a longitudinal study investigating the mathematics instructional quality of early-career elementary teachers in the United States. The automated scoring system includes a random forest classifier using the outputs of a deep neural network capable of detecting instructional activities in videos to score the mathematics instructional quality. Within this context, we present an approach to reporting the accuracy and consistency of double scoring within a *classroom observation system* when one set of scores was automated and the degree of degradation observed. This study seeks to answer the following key research questions:

1. What is the agreement between human and machine scoring? Is there a relative bias between the mean differences of human and machine scores?

2. How reliable is the machine scoring in relation to the human scoring?

3. Is the double scoring method by human and machine interchangeable to that of "gold standard" double scoring by human raters?

2 Background

2.1 Activity Detection with Deep Learning Neural Networks and Random Forests

Deep learning has become the state-of-the-art choice for various challenges including recognizing human activities in video content (Beddiar et al., 2020). A deep neural network is a hierarchical learning structure that can learn complex and abstract features of a given set of data. It is feasible to train neural networks to classify activities in videos of instruction such as the activity structure (i.e., whole group instruction, small group instruction, individual work, and transitions; Ahuja et al., 2019; Foster et al., 2024a), student and teacher behaviors (Foster et al., 2024a; Patidar et al., 2024; Sharma et al., 2021; Sun et al., 2021), and their location (Foster et al., 2024a; Patidar et al., 2024).

In this study, a deep neural network was used to detect instructional activities within video content of elementary mathematics instruction. From the output of the neural network, a random forest classifier was then used to predict the mathematics instructional quality. Random forests are a supervised machine learning algorithm that use many tree-like structures (i.e., decision trees) to make predictions or classifications (James et al., 2021). In the case of classification, a random forest selects the majority vote from decision trees.

2.2 Classroom Observation Measures for Ambitious Mathematics Instruction

There is no single conceptualization of quality mathematics instructional, although there is a fair amount of overlap in what should be regarded as high-quality instruction in mathematics (Praetorius and Charalambous, 2018; Schlesinger and Jentsch, 2016). We conceptualize high-quality mathematics instruction as teaching practices aligned with ambitious mathematics teaching (Franke et al., 2007; Lampert et al., 2013; Newmann and Associates, 1996; Thompson et al., 2013). The Mathematics Scan (M-Scan) is a classroom observation protocol for mathematics teaching aligned with ambitious mathematics instruction (Berry et al., 2013; Walkowiak et al., 2018). It is operationalized at the lesson level and has been empirically validated (Walkowiak et al., 2014). M-Scan has nine dimensions organized under four domains. For each dimension, there are indicators with descriptions for low (1-2), medium (3-5), and high (6-7) ratings.

2.3 Interrater Agreement and Reliability in Classroom Observations

A concern within classroom observation systems is whether raters can accurately and consistently apply an observational instrument. There are several approaches to reporting interrater agreement and reliability and some literature lists

these terms interchangeably (Tinsley and Weiss, 2000; White, 2018; White and Ronfeldt, 2024; Wilhelm et al., 2018). However, interrater agreement and interrater reliability are different measures. Interrater agreement indicates the extent to which different raters assign the exact same rating to each observation. Interrater reliability indicates the consistency of raters when scoring a collection. It is possible to have high interrater agreement and low interrater reliability and vice versa (Cicchetti and Feinstein, 1990). Therefore, it is important to consider the implications of both measures and their magnitude (White, 2018; Wilhelm et al., 2018).

Recent research has also brought attention to monitoring rater quality in classroom observation research (White and Ronfeldt, 2024). Typically, only a small subset of videos (i.e., reliability set) is ever double scored to monitor rater quality. However, current recommendations advise more than 20 observations to estimate rater error with 95% confidence (White and Ronfeldt, 2024). One suggestion to help guide researchers' decision making about rater errors is to use simulations. These simulations assume a level of rater error (e.g., ± 1 or 2 points) and hypothesized distribution of "true" or master scores and then examine the implications of the number of observations needed for suitable rater error rates.

2.4 Methods Comparison

This paper makes use of a technique for comparing two methods for measurement that arose out of clinical medical research (Altman and Bland, 1983, Bland and Altman, 1999, 2003, 2010). The technique is not widely used in education research, but there have been recent calls for its use (Wilhelm et al., 2018). The technique can compare one established method to another indirect or less costly alternative method. It assumes that neither method can provide a true measure and so seeks to determine how much the two methods agree and whether they are interchangeable in practice. Both methods are applied to the same observations. Then, the difference between the measures for each observation is calculated to compute the mean difference ($\bar{d}$). If the mean difference is non-zero, then this indicates there is a relative bias.

A range, in which most differences between measurements by the two methods will lie, is called the limit of agreement (LOA). This LOA can be determined using parametric and non-parametric approaches (Bland and Altman, 1999). In this study, we take a non-parametric approach as the distribution of between-method differences (i.e., difference in human ratings) is not well-known. First, we order the differences observed from least to greatest. Then, we remove 5% of the observed differences beginning with the most extreme from either end of the distribution. After removing those extreme differences, we report the LOA by finding the difference of the remaining two endpoints of the observed differences.

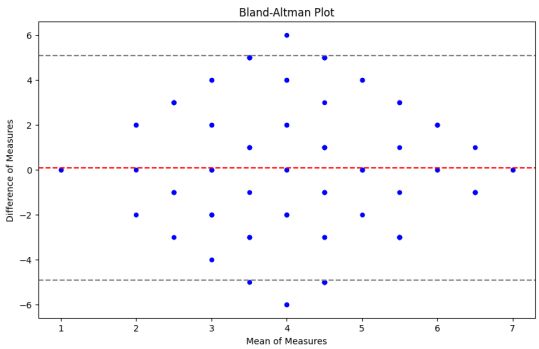

Figure 1: A Bland-Altman plot of two simulated measurements.

There is a related graphical representation, referred to as a Bland-Altman plot. It plots the average of the two methods against the difference of the two methods for each observation. Figure 1 shows a Bland-Altman plot of two simulated measurements of 100 lesson observations. The first and second measure observations range between 1 and 7. In Figure 1, we see that $\bar{d} = 0.08$ and $LOA = 10$. From this Bland-Altman plot, we interpret there is little to no relative bias when $\bar{d}$ is close to 0 and we could say that for 95% of observations, a measurement by the first approach would differ no more than ± 5 units from the second approach. If $LOA \leq 10$ is negligible in practice, then we may conclude the two methods for measuring are interchangeable.

3 Current Study

3.1 Video Data

Videos of elementary instruction used in this study were collected as part of a previous research study known as the Developing Ambitious Instruction (DAI, Youngs et al., 2022). The DAI focused on 83 beginning elementary teachers who graduated from teacher preparation programs at five universities in the United States either in 2015-16 or 2016-17. After graduation, these individuals

started teaching young children (ages 5 to 11) full-time in grades K-5 in general education settings. The DAI team observed each teacher as they taught mathematics and English language arts (ELA) at least six times each during their first two years of full-time teaching (i.e., three times each for mathematics and ELA per school year). Each video-recorded lesson was about 45 minutes in length and the current study used a total of 360 hours of video from over 400 lessons.

3.2 Scoring of Lesson Videos with M-Scan

As part of the DAI study, videos of mathematics lessons were assigned scores with M-Scan by at least one human rater. The following steps were taken to train M-Scan raters and ensure high levels of reliability. First, each rater watched three videos of elementary mathematics lessons, assigned scores for each M-Scan domain, and reviewed the master ratings and justifications. Raters then met with the master rater and watched video clips that exemplified different scores on each of the M-Scan dimensions and practiced rating two additional lessons. To determine if raters met certification requirements, raters independently coded a series of lessons without conferring with the master rater; then they met with the master rater to confer on scores. The master rater computed agreement scores (at least 80% exact or adjacent matches were required), identified items that were sources of systematic error, and looked at convergence of ratings. If a rater did not meet the 80% threshold, they were required to rate an additional two lessons. On a regular basis, the master rater conducted a meeting in which raters viewed, coded, and discussed one or two lessons from the reliability set. These meetings were used to monitor raters' ongoing performance.

3.3 Annotations of Videos

For the purpose of training a neural network, the team developed a list of 24 instructional activities for annotating the video dataset. For example, the annotation label of *Using or holding an instructional tool* was developed in reference to the M-Scan dimension Students' Use of Math Tools. Prior to annotating the video dataset, annotators went through training on how to apply the classroom-based activity labels. At the end of the training, annotators' performance was periodically monitored (Foster et al., 2024c).

3.4 Neural Network Model for Instructional Activity Detection

From our prior investigation, we found The Background Suppression network (BaS-Net, Lee et al., 2020) was advantageous for detecting activities in classroom videos (Foster et al., 2024b). The 268 hours of video recordings were used to train and test a modified BaS-Net to detect the 24 instructional activity labels, which we call BaS-Net+. In our experimental setup, training and testing sets were split 80 and 20 percent respectively. In comparison to previous reported results (Foster et al., 2024a), we restructured the testing set so that it did not feature any of the teachers from the training set. Once the neural network was trained and tested on 268 hours from DAI dataset, we then used it to detect the 24 instructional activities in an evaluation set featuring 92 additional math lessons.

3.5 Random Forest Classifiers for M-Scan Scoring

Random forest classifiers were used to predict scores for each M-Scan dimension. We developed the random forest classifiers with the package `randomForest` in R (Breiman, 2001). All 24 instructional activity labels generated by human annotations were used as initial predictors for each of the nine M-Scan dimensions scores. Each random forest included 41 decision trees with the mtry hyper-parameter set between 3 and 5 features at each step of branching.

After building the random forest classifiers, we applied them to the aggregated data in the evaluation set that was generated by BaS-Net+. We then compared the predicted score by the random forest classifiers to human scores.

3.6 Measuring Interrater Agreement and Reliability

We report agreements as ratings that agree exactly or differ by no more than 1 point (Lawlis and Lu, 1972), which we denote as p_0 and p_1. These levels of agreement are often used in practice with human raters for M-Scan (Walkowiak et al., 2018). For interrater agreement, we also report a descriptive index of agreement, developed by Tinsley and Weiss (1975), called the T-index:

$$T = \frac{N_a - Np_c}{N - Np_c} \quad (1)$$

where N_a is the number of agreements, N is the number of ratings, and p_c is the probability of chance agreement of an individual. If T is positive, then the observed agreement is greater than agreement that would occur by chance. If T is negative, then the observed agreement is less than chance agreement. When T is zero, then the observed agreement is equal to the expected chance agreement. There is also a nonparametric chi-square test of significance for the T-index (Lawlis and Lu, 1972; Tinsley and Weiss, 2000). We use T_0 and T_1 to indicate the index that corresponds to exact agreement and within one point agreement, respectively.

To report any relative bias between machine and human ratings, we report the mean difference $(\bar{d})$. A positive mean difference indicates on average the machine scored higher than the human raters while a negative value indicates on average the human scored relatively higher for a given dimension of M-Scan. We use a threshold of $\bar{d} > 0.20$ to conclude a possible relative bias between machine and human ratings.

When reporting interrater reliability for ordinal-scaled ratings, we use Finn's coefficient (r_f) and Gwet's AC_2 for their respective advantages. Both indices range between 0 and 1 with higher values indicating higher levels of consistency between raters. A nonparametric chi-square test for significance is available for r_f and AC_2. Finn's coefficient is recommended for use when the within-raters variance is highly constrained and to use a $p < 0.01$ for applied research (Tinsley and Weiss, 2000). It does not require independent subjects, which in our use case is important as some of the lessons were taught by the same teacher. Gwet's AC_2 is a generalization of Gwet's AC_1 (Gwet, 2008) and it does not assume all raters will be paired randomly for each observation.

3.7 Comparing Methods of Double Scoring

To determine two methods are interchangeable, the Bland-Altman method requires specificity beforehand as to how small the LOA should be to conclude that either method is sufficient in practice. This decision is a practical one, not a statistical decision (Bland and Altman, 1999). Practitioners should provide a strong rationale for this decision. We decided to use what has been observed in prior research with human raters as the "gold standard," although some could argue this may not be sufficient evidence (White and Ronfeldt, 2024). With expert human raters scoring with M-Scan, it was found that they agreed exactly 66.7% and within one point 97.6% of the time (Walkowiak et al., 2018). Thus, we decided to use 65% exact agreement and 95% agreement within one point. We may conclude the two methods are interchangeable if they met or exceeded each of these levels of agreement if the $LOA \leq 1$. However, before we may make such a conclusion, we must check the assumption that there is no relation between the difference between the ratings and average ratings. We use Spearman's rank correlation coefficient (ρ) to examine for any monotonic relations. We use the criteria $|\rho| > 0.30$ to conclude the possibility of any monotonic relationship.

4 Results

4.1 Research Question 1: Agreement

The exact agreement between human and machine scoring ranged between 10.9% to 58.7%. The corresponding T_0-index values are listed in Table 2. Most T_0 -index values indicated little to no agreement between the scoring except for Mathematical Accuracy, which indicated moderate agreement ($0.41 \leq T_0 < 0.60$). Allowing for one point difference, we found the extended percent agreements of human and machine ratings between 57.6% and 89.1% agreement. The corresponding T_1 -index values range between 0.31 and 0.82. These are moderate to substantial ($T_1 \geq 0.61$) levels of agreement between human and machine ratings. Almost all agreements between human and machine ratings for each dimension of M-Scan were found to be statistically significant; thus, it is highly unlikely these levels of agreement were the result of chance.

M-Scan Dimension	Interrater Agreement			
	T_0	p_0	T_1	p_1
Structure of the Lesson	0.18***	29.3%	0.70***	81.5%
Use of Representations	0.09	21.7%	0.50***	69.6%
Students' Use of Math Tools	-0.04	10.9%	0.49***	68.5%
Cognitive Demand	0.11*	23.9%	0.40***	63.0%
Math Discourse Community	0.22***	33.7%	0.50***	69.6%
Explanation and Justification	0.14**	26.1%	0.52***	70.7%
Problem Solving	0.11*	23.9%	0.31***	57.6%
Connections and Applications	0.18***	29.3%	0.57***	73.9%
Mathematical Accuracy	0.52***	58.7%	0.82***	89.1%

*$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Interrater agreements.

Next, we report if there were any relative bias between the human and machine ratings. Examining the mean differences between the paired human and machine scores (see Table 3), we found a relative bias for nearly all the M-Scan dimensions. The human raters were, on average, rating higher scores in comparison to the random forest classifiers for some of the M-Scan dimensions (e.g., Problem Solving). On other dimensions, the random forest classifiers were, on average, scoring higher than the human raters (e.g., Explanation and Justification). No systematic bias was found for the dimensions of Use of Representation when observing the mean differences between the human and machine ratings.

M-Scan Dimension	Mean Differences
	$\bar{d}$
Structure of the Lesson	0.48
Use of Representations	-0.03
Students' Use of Math Tools	-0.23
Cognitive Demand	-0.82
Math Discourse Community	-0.64
Explanation and Justification	0.73
Problem Solving	-0.99
Connections and Applications	-0.33
Mathematical Accuracy	0.30

Table 3: Mean differences between machine and human ratings.

4.2 Research Question 2: Reliability

For all dimensions of M-Scan, we found the interrater reliability between human and machine ratings to be more than substantial (> 0.600) and statistically significant ($p < 0.001$) according to Finn's reliability coefficient (r_F) and Gwet's AC_2. These interrater reliability statistics for each dimension of M-Scan are listed in Table 4.

M-Scan Dimension	Interrater Reliability	
	r_F	AC_2
Structure of the Lesson	0.812***	0.855***
Use of Representations	0.765***	0.851***
Students' Use of Math Tools	0.621***	0.757***
Cognitive Demand	0.800***	0.814***
Math Discourse Community	0.812***	0.860***
Explanation and Justification	0.722***	0.838***
Problem Solving	0.702***	0.706***
Connections and Applications	0.802***	0.877***
Mathematical Accuracy	0.891***	0.946***

*$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Interrater reliability

4.3 Research Question 3: Interchangeability

In this section, we report whether the double scoring done by human and machine is interchangeable to the "gold standard" between human raters. For our purpose, we decided if we observed at least 95% agreement within the $LOA \leq 1$ and at least 65% exact agreement between human and machine ratings, then the double scoring done by human and machine would be interchangeable with the method of two human raters. Meeting this condition would indicate that the method of rating a lesson by a human rater and machine rater agrees sufficiently in practice. Table 5 lists all the LOA for each dimension of M-Scan. Before finding the LOAs using the Bland-Altman method, we checked the assumption needed that there is no relation between the difference between the ratings and average ratings using Spearman's ρ-statistic. As shown in Table 5, all dimensions except Problem Solving did not satisfy the needed assumption; thus, these LOA should be interpreted with caution. Nevertheless, we found no evidence to suggest the method of pairing human raters with any of the random forest classifiers is interchangeable with the double scoring with two human raters. This conclusion came from the two

necessary criteria: the exact agreement was ≥65% and at least 95% agreement for a $LOA \leq 2$.

M-Scan Dimension	Limit of Agreement	Spearman's Coefficient
	LOA	ρ
Structure of the Lesson	5*	-0.37
Use of Representations	4*	-0.99
Students' Use of Math Tools	6*	-0.80
Cognitive Demand	5*	-0.81
Math Discourse Community	4*	-0.77
Explanation and Justification	5*	-0.82
Problem Solving	6	-0.24
Connections and Applications	4*	-1.0
Mathematical Accuracy	3*	-0.88

Note: (*) indicates these *LOA* interpretations should be interpreted with caution as there is an association between the mean score and scoring difference, as evidenced by corresponding value of ρ, which does not satisfy one of the criteria for use of the Bland-Altman method.

5 Discussion

Rater error is highly complex and so it is difficult to claim that raters are not significantly altering a measure such as instructional quality. Although interrater agreement and reliability provide some estimates of rater error, recent research suggests a precise measure of rater error requires more scoring occasions than what is typical (White and Ronfeldt, 2024). As a result, this means there is a significant need to double score a sizeable collection to capture a robust measure of rater error.

One potential solution to meeting this size of double scoring is to develop an automated rater. We used our study as a context to illustrate an approach for determining whether double scoring when one of the raters is an automated scoring system is interchangeable with the "gold-standard" of two human raters. We drew on classroom observation systems research and methods comparison studies.

In the context of this study, we found insufficient evidence that the method of double scoring the video by a human and machine was interchangeable with the "gold-standard" method of double scoring by two human raters. Although we found some agreement and reliability between the human and machine ratings, the current level of performance did not provide evidence for the ability to interchange the two methods as set by our outset criteria from what had previously been observed. We acknowledge decisions that we made may not be appropriate for every scoring design.

However, this study goes beyond what is typically reported in findings about the performance of automated classroom observation systems, which typically detail the association between human and machine scores. This study also examined potential impacts on scoring design decisions as they relate to automated scoring such as double scoring when one rater is an automated system. This decision could have several consequences for rater monitoring and associated time and financial costs. There is a need for evaluators of these automated systems to consider methods and frameworks for addressing this issue and others that are beyond calibration between human and machine raters (c.f., Doewes et al., 2023; Johnson et al., 2022; Rotou and Rupp, 2020; Williamson et al., 2012).

Acknowledgments

We would like to thank the members of the Artificial Intelligence for Advancing Instruction team and the Development of Ambitious Instruction team at the University of Virginia for their contributions. This work was supported by the National Science Foundation under Grant No. 2000487 and The Robertson Foundation. Opinions, findings, and conclusions in this presentation are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26.

D. G. Altman and J. M. Bland. 1983. Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(3):307–317.

Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. 2020. Vision-based human activity recognition: A survey. *Multimedia Tools and Applications*, 79(41):30509–30555.

Robert Q Berry, Sara E Rimm-Kaufman, Erin M Ottmar, Temple A Walkowiak, Eileen G Merritt, and Holly H Pinter. 2013. The Mathematics Scan (M-Scan): A measure of standards-based mathematics teaching practices.

J. M. Bland and D. G. Altman. 2003. Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology*, 22(1):85–93.

J Martin Bland and Douglas G Altman. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160.

J. Martin Bland and Douglas G. Altman. 2010. Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 47(8):931–936.

Jonathan Bostic, Kristin Lesseig, Milan Sherman, and Melissa Boston. 2021. Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education*, 24(1):5–31.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

D. V. Cicchetti and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In pages 103–113.

Jonathan K. Foster, Matthew Korban, Peter Youngs, Ginger S. Watson, and Scott T. Acton. 2024a. Automatic classification of activities in classroom videos. *Computers and Education: Artificial Intelligence*, 6:100207.

Jonathan K. Foster, Matthew Korban, Peter Youngs, Ginger S. Watson, and Scott T. Acton. 2024b. Classification of instructional activities in classroom videos using neural networks. In Xiaoming Zhai and Joseph Krajcik, editors, *Uses of Artificial Intelligence in STEM Education*, pages 439–464. Oxford University Press.

Jonathan K. Foster, Peter Youngs, Rachel van Aswegen, Samarth Singh, Ginger S. Watson, and Scott T. Acton. 2024c. Automated classification of elementary instructional activities: Analyzing the consistency of human annotations. *Journal of Learning Analytics*:1–18.

Megan L. Franke, Elham Kazemi, and Dan Battey. 2007. Mathematics teaching and classroom practice. In Frank K. Lester and National Council of Teachers of Mathematics, editors, *Second handbook of research on mathematics teaching and learning: a project of the National Council of Teachers of Mathematics*, pages 225–256. Information Age Publishing, Charlotte, NC.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Heather C. Hill, Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4):430–511.

Heather C. Hill, Charalambos Y. Charalambous, and Matthew A. Kraft. 2012. When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2):56–64.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *Artificial Intelligence in Education*, pages 60–74, Cham. Springer Nature Switzerland.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An introduction to statistical learning: With applications in R*.Springer texts in statistics. Springer, New York, NY, Second edition.

Matthew S. Johnson, Xiang Liu, and Daniel F. McCaffrey. 2022. Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. *Journal of Educational Measurement*, 59(3):338–361.

Magdalene Lampert, Megan Loef Franke, Elham Kazemi, Hala Ghousseini, Angela Chan Turrou, Heather Beasley, Adrian Cunard, and Kathleen Crowe. 2013. Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, 64(3):226–243.

G. Frank Lawlis and Elba Lu. 1972. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78(1):17–20.

Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. 2020. Background Suppression Network for weakly-supervised temporal action localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11320–11327.

Shuangshuang Liu, Courtney A. Bell, Nathan D. Jones, and Daniel F. McCaffrey. 2019. Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1):61–95.

Fred M. Newmann and Associates. 1996. *Authentic achievement: Restructuring schools for intellectual quality*.The Jossey-Bass education series. Jossey-Bass, San Francisco.

Prasoon Patidar, Tricia Ngoon, Neeharika Vogety, Nikhil Behari, Chris Harrison, John Zimmerman, Amy Ogan, and Yuvraj Agarwal. 2024. Edulyze: Learning analytics for real-world classrooms at scale. *Journal of Learning Analytics*, 11(2):297–313.

Anna-Katharina Praetorius and Charalambos Y. Charalambous. 2018. Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, 50(3):535–553.

Ourania Rotou and André A. Rupp. 2020. Evaluations of automated scoring systems in practice. *ETS Research Report Series*, 2020(1):1–18.

André A. Rupp. 2018. Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, 31(3):191–214.

Lena Schlesinger and Armin Jentsch. 2016. Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, 48(1):29–40.

Vijeta Sharma, Manjari Gupta, Ajai Kumar, and Deepti Mishra. 2021. EduNet: A new video dataset for understanding human activity in the classroom environment. *Sensors (Basel, Switzerland)*, 21(17):5699.

Bo Sun, Yong Wu, Kaijie Zhao, Jun He, Lejun Yu, Huanqing Yan, and Ao Luo. 2021. Student Class Behavior Dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Computing and Applications*, 33(14):8335–8354.

Jessica Thompson, Mark Windschitl, and Melissa Braaten. 2013. Developing a theory of ambitious early-career teacher practice. *American Educational Research Journal*, 50(3):574–615.

Howard E. Tinsley and David J. Weiss. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22(4):358–376.

Howard E. Tinsley and David J. Weiss. 2000. Interrater reliability and agreement. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, pages 95–124. Elsevier.

Temple A. Walkowiak, Robert Q. Berry, J. Patrick Meyer, Sara E. Rimm-Kaufman, and Erin R. Ottmar. 2014. Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1):109–128.

Temple A. Walkowiak, Robert Q. Berry, Holly H. Pinter, and Erik D. Jacobson. 2018. Utilizing the M-Scan to measure standards-based mathematics teaching practices: affordances and limitations. *ZDM*, 50(3):461–474.

Rose E. Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. arXiv:2306.03090 [cs].

Mark C. White. 2018. Rater performance standards for classroom observation instruments. *Educational Researcher*, 47(8):492–501.

Mark White and Matt Ronfeldt. 2024. Monitoring rater quality in observational systems: Issues due to unreliable estimates of rater quality. *Educational Assessment*, 29(2):124–146.

Anne Garrison Wilhelm, Amy Gillespie Rouse, and Francesca Jones. 2018. Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability. *Practical Assessment, Research, and Evaluation*, 23(4):1–16.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Peter Youngs, Lauren Molloy Elreda, Dorothea Anagnostopoulos, Julie Cohen, Corey Drake, and Spyros Konstantopoulos. 2022. The development of ambitious instruction: How beginning elementary teachers' preparation experiences are associated with their mathematics and English language arts instructional practices. *Teaching and Teacher Education*, 110:103576.