"All that glitters": Techniques for Evaluations with Unreliable Model and Human Annotations

Anonymous ACL submission

Abstract

"Gold" and "ground truth" human-mediated labels have error. This error can escape commonly reported metrics of label quality or obscure questions of accuracy, bias, fairness, and usefulness during model evaluation. This study demonstrates methods for answering such questions even in the context of very low reliabilities from expert humans. We analyze human labels, GPT model ratings, and transformer encoder model ratings of the quality of classroom teaching from two LLM architecture families–encoders and GPT decoders. First, we demonstrate that using standard metrics in the presence of poor labels can mask both label and model quality: the encoder family of models achieve state-of-the-art, even "super-human", results across all classroom annotation tasks. But not all these results stand after using more rigorous evaluation measures, revealing spurious correlations and nonrandom racial biases across models and humans. Finally, we expand these methods to estimate how model use would change to human label quality if models were used in a human-in-the-loop context, finding that the variance captured in GPT model labels would worsen reliabilities.

1 Introduction

Human mediated labels always have an unknown amount of error. In machine learning practice, this uncertainty is typically ignored during supervised model development and evaluation. Imperfect labels are treated as "gold" or "ground truth". This may be due in part to measures of accuracy being the most preferred methods of assessing and benchmarking model performance (Ribeiro et al., 2020; Kiela et al., 2021). Standard correlations or inter-rater agreement metrics are common measures of the quality of dataset annotations to justify the further use of the labels with supervised learning and model performance evaluation. When working with low/unknown reliability annotations, more often found in tasks requiring complex expert judgment, what methods can be used to improve evaluation under label uncertainty?

More nuanced approaches can account for sources of variation and types of uncertainty. This study focuses on model–and human–label evaluation, demonstrating latent variable and random effect techniques that could be used to also inform dataset curation and allow more intentional model training (e.g., (Gordon et al., 2022)). The analyses presented in this study are motivated by issues of model interpretability, fairness, and usefulness.

Better understanding human label behaviors is key to training and evaluating models (Webson et al., 2023; Webson and Pavlick, 2022). To illustrate more rigorous methods of working with label and prediction uncertainty, this study examines an extreme use case where expert annotations are highly unreliable yet simultaneously used in high-stakes decisions: automated rating of the quality of classroom teaching. Contributions of this work include: (a) novel use of psychometric methods for evaluating LLMs on human tasks, (b) detection of spurious correlations in model outputs via disattenuating low human-model correlations, (c) comparison of individual model behaviors with individual human annotator behaviors by disentangling rater-level effects, (d) measurement of model fairness and bias in the presence of low label reliabilities, and (e) application of Design Studies (D-studies) from Generalizability Theory (G-theory) for estimating impacts of human-in-the-loop (HIL) model use on human label quality. Brief introductions to various psychometric techniques will be provided and illustrated via the case study, with interpretation of their limitations and recommendations for future research.

1.1 Case Study: Rating Teaching Quality

This classification task may seem deceptively simple: using a rubric, provide a rating for the quality of instruction of an elementary school math classroom. Such ratings are given to all US K12 public education teachers for both formative educator development feedback and as high-stakes teacher evaluations. Despite their ubiquity, these ratings, even when conducted by experts, are unreliable, similar to the poor reliability of other K12 education labels (Jurenka et al., 2024; Wang and Demszky,

2023; Whitehill and LoCasale-Crouch, 2024; Ho and Kane, 2013; Kane et al., 2015; Kane and Staiger, 2012; Glaese et al., 2022) that have limited the rigor of education research (Slavin, 2002; Klahr, 2013).

Using similar methods, Ho and Kane estimated that increasing the number of human classroom observers can improve the reliability of ratings assigned. In their major work on the topic, they measure conditions under which the use of additional human raters can increase the reliability of this resource- and time-intensive task (Kane and Staiger, 2012; Whitehurst et al., 2014). Similarly, to answer whether automated ratings can similarly improve human annotations, we need to understand the extent to which models' added contributions would result in the same kinds of benefits expected from humans. Are the behaviors of models similar enough to those of humans that their contributions can help in lieu of the humans recommended by Ho and Kane?

To explore this and to support more generalizable understanding of these contributions, the case study analyzes two types of raters, human and model. The model raters are comprised of two families: three models in the "GPT" family of autoregressive in-context learners from (Wang and Demszky, 2023) (using gpt-3.5-turbo) whose siblings differ by prompt engineering strategies, and five models in an "Encoder" family built for this study whose siblings differ in embeddings and a few model training hyperparameters. Quality of ratings will be examined between and within families and individual raters.

This work strengthens the argument that only using simple inter-rater reliability metrics to understand the quality of labels may be masking the limitations of the labeling criteria (Hill et al., 2012b). It also illustrates how more robust evaluation techniques can yield information in the presence of noisy labels and seemingly inconclusive results.

2 Related Work

2.1 Annotation Quality and Bias

Accuracy, based on "gold" or "ground truth" labels, is the primary type metric by which LLMs are evaluated (Ribeiro et al., 2020; Kiela et al., 2021). For expediency of development, data scientists need to assume data labels are reliable, accurate, and end-task aligned for intended real-world use cases, (Bejar et al., 2006; Messick, 1998), even in scenarios where these assumptions could be detrimental (e.g., performing complex high-stakes tasks, reducing discriminatory biases found in data (Field et al., 2021) that are immutably historical by definition of their creation, etc.), which is especially true of autoregressive models, whose labels are Internet text and which contain harmful biases (Hofmann et al., 2024a,b). Assessing the accuracy and reliability of idiosyncratically human annotated "ground truth" can be difficult (Eckes and Jin; Wind and Guo, 2019; Wind, 2019; Abercrombie et al., 2023; Baan et al., 2024, 2022; Waseem, 2016; Kazai et al., 2013; Hosseiny Marani et al., 2022), a challenge that is exacerbated when label uncertainty is underexamined or underreported. Limited transparency around label quality makes it more challenging to measure biases, interpret model findings, assess individual fairness, and establish real-world validity.

Powerful and provocative research has begun to address the limitations of accuracy-only evaluations and propose more fair and responsible solutions (Hardt et al., 2016; Dwork et al., 2012; Kasy and Abebe, 2021; Song et al., 2020; Zhao and Ermon, 2021; Corbett-Davies et al., 2023; Pleiss et al., 2017; Zemel et al., 2013), including techniques for addressing when labels lead to undesirable model behaviors (Ding et al., 2022; Hebert-Johnson et al., 2018). This paper offers several ways to quantify these issues and improve interpretability and explainability (Adebayo et al., 2020; Lundberg and Lee, 2017; Rudin, 2019; Kim et al., 2018).

2.2 Teacher Development and Evaluation

School leaders working with teachers to improve the quality of instruction typically: evaluate the teacher's proficiency in a range of competencies (typically measured during in-class observation and evaluation on a teaching rubric; (Aguilar, 2013; Bambrick-Santoyo, 2016, 2018)), then determine which competencies are most important to improve first (i.e., which change will have the biggest impact on student learning), and then provide supportive feedback and coaching. This paper focuses on the first step of evaluating teacher proficiency, which is often time-consuming and produces ratings (labels) that are unreliable (Kane and Staiger, 2012; Blazar, 2018; Kane et al., 2013; Casabianca et al., 2013). Without accurate classifications, it is challenging for practitioners to prioritize instructional needs and aligned practices from among the many elements of good teaching (Saphier et al., 2008; Darling-Hammond, 2014; Hammond, 2015; Lemov and Atkins, 2015; Lemov, 2021; Liljedahl et al., 2021; Darling-Hammond et al., 2020; Schwartz et al., 2016) and for edtech researchers to quantify good teaching (Jurenka et al., 2024).

Thus, this work provides a bridge to research seeking to improve teaching quality by providing feedback to teachers on various instructional techniques (Samei et al., 2014; Donnelly et al., 2017; Kelly et al., 2018; Demszky et al., 2021; Suresh et al., 2022; Jacobs et al., 2022; Alic et al., 2022; Demszky and Liu, 2023; Demszky et al., 2024, 2023). These studies identify linguistic features correlated with an aspect of good teaching, but may optimistically overgeneralize the usefulness, efficacy, and universality of their solution, providing specific prescriptions

without diagnosis. Matching these models with the specific needs of teachers will help provide a more individualized approach to teacher development, one based on understanding instructional needs and then providing corresponding supports.

Only two recent studies have sought to use LLMs to provide ratings of classroom instruction (via classroom transcripts) using authentic rating rubrics. Whitehill and LoCasale-Crouch (2024) use a mix of zero-shot and bag-of-words model configurations to provide scores to instructional domains for Pre-Kindergarden classrooms using a private dataset, commenting on their highest Pearson R statistic of 72 experiments ($R = 0.48$) that it "approaches human inter-rater reliability". Wang and Demszky (2023) use publicly available datasets to evaluate zero-shot prompting, which data and outputs will be discussed below.

3 Data

The data used in this study and in (Wang and Demszky, 2023) are from the National Center for Teacher Effectiveness (NCTE) Main Study (Kane et al., 2015), which contains three years of data collection and observations of math instruction in approximately fifty schools and three-hundred (4th and 5th grade) mathematics classrooms across four school districts in the United States, including expert human ratings of individual video-captured classroom lessons across two observation instruments (Bacher-Hicks et al., 2017, 2019): the CLASS framework (12 items) (Pianta et al., 2008) for general instructional practice and the content-specific Mathematical Quality of Instruction (MQI; 13 items) (Hill et al., 2008), together yielding over 400,000 distinct human rating labels assigned, the distributions of which are in Figure 3. Each instrument item is intended to measure a different aspect of teaching quality. Like all human mediated labels,[1] an individual classroom observation rating requires at a minimum three facets: (1) task rating criteria, (2) raters/labelers, (3) observations to be classified. As tasks increase in complexity, three facets contribute more error to estimates.

3.1 Rating Criteria: MQI Rubric

Just as all raters contribute uncertainty to a system, so too do the measurement instruments. Ambiguity uncertainty is introduced when an instrument, instruction, or criteria for a task has language that could lead to two equally-expert raters to different results, ceteris paribus. The 13 MQI items within the dataset have at least two raters per classroom observation. While both humans and Encoders evaluated all items, the this paper will focus on the 4 of the 13 MQI items evaluated in (Wang and Demszky, 2023). These four ternary items are teacher explanations (EXPL), remediation of student errors (REMED), student questioning and reasoning (SMQR), and imprecision in mathematical language (LANGIMP).[2] Analyses for all other items are in the appendices. Prior studies have explored the reliability of MQI instrument ratings generally (Kane and Staiger, 2012; Mantzicopoulos et al., 2018; Hill et al., 2012b; Kane et al., 2015; Ji, 2023); this study confirms previous findings via reproduced reliability metrics in Section 3.2, which correspond to the NCTE Study, Appendix Section 2).

3.2 Human Expert Raters

Human rater information for both the MQI and CLASS instruments can be found in the Appendix of the *DS0 Study-Level Files* from the NCTE Main study. MQI raters in particular were recruited from a separate pool of applicants based on their background in mathematics and through contacting colleagues in mathematics departments (Hill et al., 2012a; Blazar et al., 2017) and then passed certification exams to score the MQI, and attended biweekly calibration meetings to ensure standardization of scoring procedures.

3.3 Classroom Observations

63 human raters watched videos and provided ratings at regular intervals across all items in the MQI. Transcripts of these same videos (Demszky and Hill, 2022) are used by LLMs for the same task, where the class discourse is equipartitioned across utterances (GPT family models) or words (Encoder family models) by the total number of classroom segments to align the text to the human labels in the absence of timestamps. Data from the NCTE Main study (Kane et al., 2015) [3] and for the associated transcripts (Demszky and Hill, 2023)[4] are available online.

4 Model Families and Model Rater Data

GPT Models The GPT model family from Wang and Demszky (2023)[5] have 7,660 ratings for 223 different teachers. The family consists of three models differing in prompt engineering methods, and brief summary of those differences is in Table 3.

[1] Label(er), rate(r), annotat(ion/or), and score(r) will be used interchangeably for these classification tasks, as terminology varies multidisciplinarily.

[2] LANGIMP is reverse-coded so higher scores are better and has interesting self-referentiality vis-à-vis instrument uncertainty that is worth noting, but out of scope for the current study.

[3] https://www.icpsr.umich.edu/web/ICPSR/studies/36095/datadocumentation

[4] https://github.com/ddemszky/classroom-transcript-analysis

[5] https://github.com/rosewang2008/zero-shot-teacher-feedback/

GPT models were evaluated on curated selections of classroom text with the least transcriptorial noise (i.e., minimizing instances of `[inaudible]`), and were edited to indicate whether the speakers were teachers or students.

Encoder Models Encoder family models are custom transformer encoders trained on the NCTE classroom transcripts. They use fixed-parameter pretrained sentence embeddings, differing in these and in training hyperparamters, thereby exploiting LLM sensitivites to pretraining regimes (D'Amour et al., 2020; McCoy et al., 2023). A quick summary of differences is in Table 2 and more training details can be found in Appendix C. In contrast to the GPT models, the only text preprocessing used with the Encoders simply replaced all transcription notes with `[inaudible]` to mimic the uncertainty in live audio transcription, and no edits to indicate speakership were included. For the Encoder models, all model outputs[6] used in evaluations and reporting in this study were conducted with a lesson-level-stratified held-out test set that was not used during model development. Encoder models were trained a single GPU in Google Colab.

5 Evaluation Methods

Typical **reliability** metrics (see Section 5.1) provide a backdrop of descriptives that can flag issues of low quality labels. Measures of statistical **dependability** can be used for generalizing label conclusions and identifying spurious correlations (see Section 5.3), a part of improving **accuracy**. Methods for disentangling human and model label **biases** (see Section 5.4) are first demonstrated and then extended to estimate **fairness** across racial lines in Section 5.5. **Usefulness**, as measured by the amount of rating reliability improvement a model can provide to a human rater in human-in-the-loop contexts, including associated cost savings in human time (for encoder models) are in Section 5.6.

5.1 Reliability and Agreement Metrics

Baseline Human Metrics Full reproductions of all reliability metrics and calculation processes exactly as described in the NCTE Main Study Appendix Section 2 were conducted. (Kane et al., 2015). Following their same procedures, replicated calculations were extended to the model families, replacing a human rater score with a *specified* or *random* model for evaluations of *individual* models and *model families*, respectively. Reproduced human results and model results, including additional metrics in this section, are fully reported in Appendix E.1.

Commonly Used Metrics The results also include three additional correlation and reliability metrics: Quadratic Weighted Kappa (QWK) typically used in ordinal classification tasks to penalize distance quadratically (squared error) while accounting for categorical agreement by chance (e.g., (Shermis, 2014; Hardy, 2021; Wang and Demszky, 2023)), Pearson correlation ρ, (e.g., (Whitehill and LoCasale-Crouch, 2024)) Spearman correlation r_s (e.g., (Wang and Demszky, 2023)), and Kendall correlation τ (e.g., (Liu et al., 2023b)). Table 4 has a full panel of lesson segment-level inter-rater reliability metrics for each MQI and CLASS item. All these common label metrics for the four items in this study are in Panel (b) in Figure 1.

5.2 Generalizable Reliability

Generalizability Study (g-study) (Brennan, 2001a, 2013, 2001b; Hill et al., 2012b) designs utilize random effect estimates across possible configurations of different sources of variance to quantify how generalizable labels are by estimating the extent to which they would persist if sources of variation changed (e.g., same teacher, different day; same lesson, different rater; human rater vs model rater; etc.). $\mathbf{E}\rho^2$ is a measure of the relative *generalizability* of a rating (i.e., is rating order preserved), and Φ, accounting for absolute error, is a measure of label *dependability*: how likely specific ratings would be numerically the same with different sources of variation. These two reliability-like estimates can help quantify how "golden" labels are.

The multifaceted g-study design used to estimate the how much variation (ν) in individual teachers' instructional quality, i, contributed to a rating label, X, annotated for a section of a lesson, s, during an observation, o, on rubric item j by rater r is known as a Item-by-Rater-by-Segment-within-Observation-within-Individual Teacher design: $J \times R \times (S : O : I)$. Simplifying the expression by conducting item-level g-studies, the resulting $R \times (S : O : I)$ design, using nested random effects notation, is:

$$X^{(j)}_{s:o:ir} = \mu + \nu_i + \nu_{o:i} + \nu_{s:o:i} + \nu_{ir} + \nu_r + \nu_{s:o:ir} \tag{1}$$

where j indicates the item index.[7] Code for the model specification is in Appendix H.2. Then, $\mathbf{E}\rho^2$ and Φ (illustrated by Equation 2) are easily estimated from the random effects for raters in rater family $\mathbb{F}$:

$$\Phi^{(j)}_{\mathbb{F}} = \frac{\nu_{ij}}{\nu_{ij} + \nu_{s:o:ij} + \nu_{irj} + \nu_{rj} + \nu_{s:o:irj}}, \tag{2}$$

[6] `github-to-be-added-after-anonymity`

[7] For the estimates in Fig. 1 (c), for depedendability metrics of Section 5.3, and for comparability with human baselines(Hill et al., 2012b; Kane et al., 2015; Ho and Kane, 2013; Kane and Staiger, 2012), a simplified model, an item-level $R \times (O : I)$ design, was conducted for the human expert rater family with results in Appendix H.1. The full model structures of Eq. 1 and 2 are used for Section 5.6.

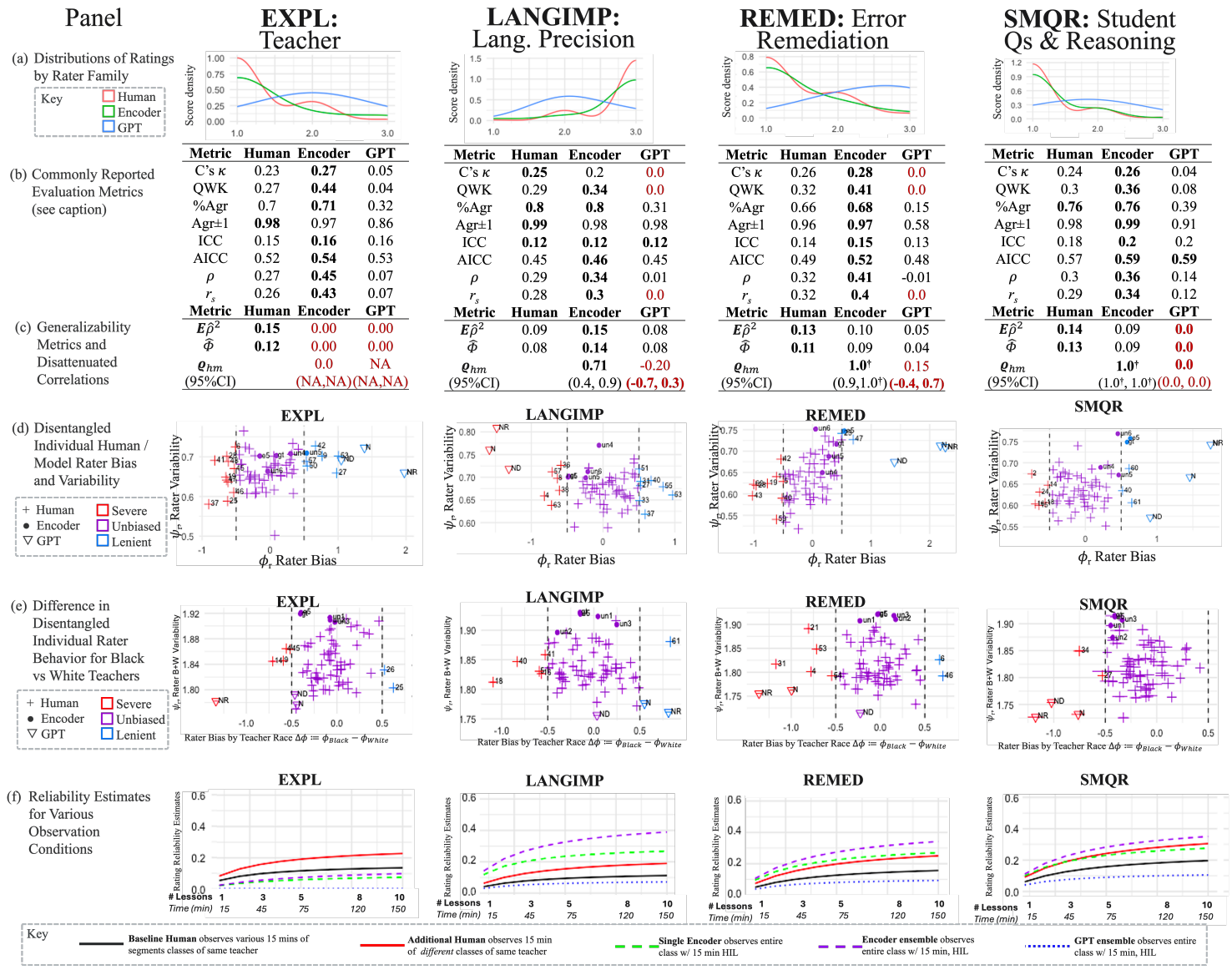

Figure 1: **Section 5 Study Method Results** for four MQI Items across Human, Encoder, and GPT rater families. **(a) Distributions**. Score distributions by rater type. **(b) Reliabilities**. Inter-rater reliability metrics introduced in Section 5.1. **C's** κ: Cohen's κ; **QWK**: Quadratic Weighted Kappa; **%Agr**: percent exact agreement; **%Agr±1**: percent agreement within 1 category; **ICC**: intraclass correlation; **AICC**: adjusted intraclass correlation; ρ: Pearson's correlation; $\mathbf{r_s}$: Spearman's rank correlation; Bold format is highest value for a given metric. **(c) Generalizability Measures and Spurious Correlation Detection**. Section 5.2: generalizability coefficient $\mathbf{E}\rho^2$ and dependability measure Φ. Section 5.3: $\varrho_{\mathbb{hm}}$ is the disattenuated correlation. Red font indicates correlation was spurious or incalculable due to low reliabilities. **(d) Disentangled Rater Bias**. Section 5.4: standardized rater bias ϕ_{jr} (x axis) and rater variability/consistency, ψ_{jr} (y axis) from Equation 6, η_j-centered. Each point represents an individual human or model rater. More severe raters are left, more lenient right. **(e) Fairness across Racial Lines**. Section 5.5: Standardized difference in rater bias ϕ_r (x axis) and rater combined variability/consistency, ψ_r, (y axis) across Black teachers and White teachers. Leftward values are more severe towards Black teachers, rightward are more lenient. Any horizontal bar present with a marker represents 95% CI for bias. **(f) Estimated Improvements to Reliability**. Section 5.6: Expected changes to rating reliability are estimated improvements to quality (via reliability) of classroom ratings for various contexts. The single individual human baseline (black) estimates reliability improvements by visiting the same class the x axis represents the number of different 15 min. classroom observations of the same teacher. The red line is estimate of having a *different* human observer conduct observations as described. By contrast, for the model raters–single Encoder (green), Encoder ensemble (average of 3 encoders) (Red), and GPT ensemble (average of 3 GPT prompt engineered models)–the x-axis for models is the number of full classroom observations conducted where the human (black) observes at least 15 minutes (in-the-loop) of the same classroom (models observe the entire class period).

$\forall r \in \mathbb{F}$. These estimates can be found in panel (c) of Figure 1.

5.3 Accuracy and Spurious Correlations

Dependability and generalizability do not guarantee accuracy, but even at these very low levels, they can be used in indirect tests of validity to see whether correlations between humans and models are low because of measurement error, such as poor item construction, or because the two sets are really uncorrelated. If an individual teacher's latent instructional ability θ_i is about the same from lesson to lesson with the same students, we can correct for measurement error by correlating human ($\mathbb{h}$) and model ($\mathbb{m}$) family ratings for *different lessons* coming from the *same teacher* and disattenuate the correlations by each rater family's, $\mathbb{F}$, label dependability for a given item, $\hat{\Phi}^{(j)}_{\mathbb{F}}$. The disattenuated correlation, $\varrho^{(j)}_{\mathbb{h}\mathbb{m}}$, between humans and a family of models for item, j, can be estimated:

$$\varrho^{(j)}_{\mathbb{h}\mathbb{m}} = \frac{\mathrm{Corr}[\tilde{\mathcal{X}}_{\mathbb{h}}(i, \mathfrak{L}, j, r_{\mathbb{h}}), \tilde{\mathcal{X}}_{\mathbb{m}}(i, \neg\mathfrak{L}, j, r_{\mathbb{m}})]}{\sqrt{\hat{\Phi}^{(j)}_{\mathbb{h}}\,\hat{\Phi}^{(j)}_{\mathbb{m}}}} \quad (3)$$

where $\tilde{\mathcal{X}}_{\mathbb{F}}$ is score retrieval function for individual teacher i on j by a random member r of rater family $\mathbb{F}$ in relation to some observed lesson $\mathfrak{L}$ with family label dependability, $\hat{\Phi}^{(j)}_{\mathbb{F}}$ defined in Equation 2. Figure 1 panel (b) has the disattenuated correlations and their respective 95% confidence intervals, calculated at $\alpha = 0.05$ using empirical confidence scaling methods defined by (Charles, 2005).

Disattenuated correlations are not directly comparable[8] to the measures of correlation in Section 5.1. However, failure of disattenuation to identify viable human-model correlations for items that previously such showed correlated relationships in Section 5.1 suggests the prior correlations may be spurious. Disattenuation does not change the low reliability across items nor the quality of the measurement, but it offers indirect evidence for discerning model predictive validity by quantifying the how changes in the underlying construct result in changes in the same direction for both human and model.

5.4 Disentangling Sources of Bias

Rater biases in complex tasks are usually not directly measurable, but we can estimate latent constructs that quantify the effects of individual raters' behaviors using methods commonly used to estimate latent attributes of rubric items (e.g., item difficulty) and latent attributes individuals (e.g., ability) throughout Item Response Theory (IRT). If the data had no variation due to raters (e.g., $\hat{\Phi}^{(j)}_{\mathbb{h}} = 1.0$), various polytomous IRT methods could help estimate "true scores"/"gold" labels (χ_{ij}) during classroom observations, teacher instructional abilities (θ_i), and the various individual item effects. For tasks with human-mediated labels, human raters introduce additional sources of measurement error for each classification and the data may include multiple measures from multiple raters for a single observation (leading to an accumulation of information at overlap observation points). To address this, hierarchical rater modeling (HRM) (Patz et al., 2002; Decarlo, 2003; DeCarlo et al., 2011) combines an IRT model with a first stage estimation defined by a signal detection theory (SDT) relationship. The latter asks the question, "given the presence of the 'true' score, can a rater detect it?" as the former asks, "given the inputs, can we estimate the 'true' score accounting for differences in the tasks used to measure it?". The hierarchical structure addresses the problem of accumulation of information in the estimates. HRMs consist of three components:

$$\text{HRM} \begin{cases} \boldsymbol{\theta}_i \sim \mathrm{MVN}(\mathbf{0}_{M\times 1}, \mathbf{I}_{M\times M}), \\ \xi_{oij} \sim \textbf{IRT model}\text{: Equation 5} \\ X_{soijr} \sim \textbf{SDT model}\text{: Equation 6} \end{cases} \quad (4)$$

where an **IRT model** estimates the "gold" label score ξ_{soij} for a given item at a particular time segment s in teacher i's o-th observed lesson for item j, which arises from i's M-dimensionally distributed **latent instructional ability**/needs ($\boldsymbol{\theta}_i$), and a **Signal Detection Theory (SDT) model** component disentangles individual rater biases from each recorded score, X_{soijr}, by quantifying the latent attributes that mediate whether rater r correctly detects the true score, i.e., $p_{\xi kr} = P\left[X_{soijr} = k \mid \xi_{oij} = \xi\ \right]$.

The IRT component of Equation 4 estimating the the true scores based on rubric item- and teacher-specific parameters is a K_j-category multidimensional generalized partial credit model (MGPCM) (Muraki, 1992; Adams et al., 1997; Cui et al., 2024; Casabianca, 2021). Distributional challenges of negatively worded items can be addressed through a multidimensional parameterization of the underlying latent teacher instructional abilities, with between-item dimensionality confirmatorily defined by the factors in (Blazar et al., 2017). The MGPCM item discrimination parameters, $\boldsymbol{\alpha}_j = \alpha_{jm}$, a vector of dimension-specific traits $\boldsymbol{\theta}_i = \theta_{im}$ are separated for $m \in M$ latent dimensions, and parameters for item difficulties γ_{jk} exist for each possible score category k in item j, $P\left[\xi_{oij} = \xi \mid \boldsymbol{\theta'}_i,\ \boldsymbol{\alpha}_j\ ,\ \gamma_{j\xi}, o\right] =$:

$$\frac{\exp\left\{(k-1)\boldsymbol{\alpha}_j\boldsymbol{\theta'}_i - \sum_{k=1}^{k}\gamma_{jk}\right\}}{\sum_{h=1}^{K_j}\exp\left\{(k-1)\boldsymbol{\alpha}_j\boldsymbol{\theta'}_i - \sum_{k=1}^{h}\gamma_{jk}\right\}}, \quad (5)$$

[8] For example, reported disattenuated correlations of 1.0 do not mean perfect correlation: it generally means that measurement error is not randomly distributed.

where $i = 1, ..., N$ lessons observed, $j = 1, ..., J$ items, $r = 1, ..., R$ raters, and $k = 1, ..., K$ possible scores.

As parameterized by (Patz et al., 2002), the base-level SDT model of the HRM represents the measurement error induced by rater r whose ability to "detect" the true score changes according to an individual rater's item-specific biases, ϕ_{jr} and variabilities, ψ_{jr} , on the x and y axes of Figure 1:

$$p_{\xi kr} \propto \exp\left\{-\frac{1}{2\psi_{jr}^2}\left[k - (\xi \ + \ \phi_{jr})\right]^2\right\} \quad (6)$$

where $\boldsymbol{\phi}_{jr} = \mathbf{Y}_{jr}\eta$ is a linear model for rating bias for items and with design matrix $\mathbf{Y}_{jr}$ of dimensions $(RJ) \times (R+J)$ and $\eta = (\phi_1, ..., \phi_R, \eta_1, ...\eta_J)^T$ for R raters and J items, as parameterized in (Mariano and Junker, 2007). Correspondingly, we update $\ln\psi_{jr}^2 = \mathbf{Y}_{jr}(\ln\tau^2)$ where $\ln\tau^2 = (\ln\psi_1^2, ..., \ln\psi_R^2, \ln\tau_1^2, ..., \ln\tau_J^2)^T$. The complete rater estimates from these models are displayed in Figure 6. The Bayesian estimates were calculated via Markov-chain Monte Carlo (MCMC) simulation using Gibbs sampling across four chains using JAGS (Plummer, 2003) in R using very weakly-informative priors, converging with $\hat{R} < 1.1$ for each parameter.

5.5 Measuring Fairness across Racial Lines

Disentangling individual rater biases further, across sensitive attributes, can provide a measure of fairness for labels and identify raters (human or model) that display discriminatory biases. Variables representing a sensitive attribute, ς (e.g., race/ethnicity, gender, age, etc.) should be independent of observed score X_{soijr} given the true score ξ_{soij} if ratings are fair: $X \perp \varsigma \Rightarrow P_{\varsigma=a}(X_{jr}|\xi_j) = P_{\varsigma=b}(X_{jr}|\xi_j), \forall a, b$. In the notation used for disentangling rater effects, there should be no difference in variation in scoring from rater r on item j is fair with respect to attribute ς given $\varsigma \perp \xi$:

$$P[X_{soijr}|\xi_{soij}, r, j, \varsigma_i] = P[X_{soijr}|\xi_{soij}, r, j] \quad (7)$$

To measure a rater's item-level fairness with respect to some sensitive teacher attribute, ς, the rater parameter vectors are easily updated where $\phi_{jr\varsigma} = \mathbf{Y}_{jr\varsigma}\eta$ is now a linear model for rating bias for items and with $\mathbf{Y}_{jr\varsigma}$ is a design matrix of dimensions $(RJ\Sigma) \times (R+J+\Sigma)$ and $\Sigma = \{B, W\}$ for Black and White self-identified teachers respectively. In this case, where $\varsigma_i \in \{B, W\}$, we can update the vector explicitly to illustrate those values $\eta = (\phi_{1_B}, \ldots, \phi_{R_B}, \phi_{1_W}, \ldots, \phi_{R_W}, \eta_{1_B}, ...\eta_{J_B}, \eta_{1_W}, ...\eta_{J_W})^T$ for R raters, J items, , and , $\ln\psi_{jr\varsigma}^2 = \mathbf{Y}_{jr\varsigma}(\ln\tau^2)$ is similarly updated such that $\ln\tau^2 = (\ln\psi_1^2, \ldots, \ln\psi_R^2, \ln\tau_1^2, ..., \ln\tau_J^2, \tau_B^2, \tau_W^2)^T$.

Less than 1% of parameter estimates had $\hat{R} \geq 1.1$, whose differences in posterior distributions have no material effect on results or discussion; all rater-item-specific 95% credible intervals for biases are represented as horizontal lines in Figure 1, in panel (e). Appendix G has full JAGS code used for the formula specification for all items and dimensions, including initial value parameters.

5.6 Estimating Real-world Helpfulness

A Decision Study (D-study) estimates how reliabilities of ratings could improve by adjusting measured facets of variation, much like Ho and Kane did to motivate the case study. Its associated estimates are in Figure 1 panel (f). To estimate the reliability in a human-in-the-loop scenario, multiple g-studies and d-studies would need to be constructed to combine the variance contributions across a set rater families, $\mathbb{F}$. For this work, only two different types of families are consider in each d-study, and one of them will always be human, as automated rating models, even high-performing Encoders, are not yet ready to produce ratings independent from human confirmation. For a human-in-the-loop decision study, $\mathbb{F}$ would consist of families that have humans only and models only, and a combined human-model family. For a $(S : O : i) \times R$ study on item j estimated reliability of ratings provided to a teacher i, $\tilde{\rho_{ij}}$, in the joined "universe" $\mathbb{F}'$ are a function of the absolute error, $\sigma_j^2(\Delta)$, as it changes under different settings, $\mathbf{K}$:

$$\widetilde{\rho_{ij,\mathbf{K}}^{(\mathbb{F}')}} = \frac{\sum_\pi^{\mathbb{F}} \sigma_j^2(i_\pi)}{\sum_\pi^{\mathbb{F}} \sigma_j^2(i_\pi) + \sigma_j^2(\Delta_\pi)} \quad (8)$$

where the summations in Equation 8 combines the variation across the familial "universes", indexed by π, of different rater families in $\mathbb{F}$ and $\sigma_j^2(i_\pi)$ and $\sigma_j^2(\Delta_\pi)$ represents the "universe" variability for teacher i and the absolute error, respectively, at the teacher-year-level (i) across the combined parameterization set $\mathbf{K}$. Structurally, Equation 8 shares similarities with the two-stage ICC calculation of Eq. 12. The absolute error for an indexed rater family (π) across any permutation of decision values in this study:

$$\sigma_j^2(\Delta_\pi) = \frac{\sigma^2(r_\pi)}{n'_{r_\pi}} + \frac{\sigma^2(o_\pi : i)}{n'_{o_\pi}} + \frac{\sigma^2(r_\pi i)}{n'_{r_\pi}} \quad (9)$$

$$+ \frac{\sigma^2(s_\pi : o_\pi : i)}{n'_{s_\pi} n'_{o_\pi}} + \frac{\sigma^2(s_\pi : o_\pi : ir_\pi)}{n'_{s_\pi} n'_{o_\pi} n'_{r_\pi}} \quad (10)$$

where the decision values vary across design facets and whose contribution is weighted by the combined count n'_k of a given facet k for ratings generated only by the family indexed by π, n_{k_π} and those facets, if any, shared between families, $n_{k_{\mathbb{F}'}}$: $n'_{k_\pi} = n_{k_\pi} + n_{k_{\mathbb{F}'}} \forall k \in \{s, o, r\}, n_{r_{\mathbb{F}'}} = 0$. These distinct sets of parameter values for each design study are represented in Equation 8. For human-in-the-loop *only* use cases, $n_{k_{\mathbb{F}'}}$ represents those sources of variation that are shared between rater families, and for a model family $\pi = \mathbb{m}$, where there would be

no observations made by a model without a human, the model would not have any independent observations $n_{o_{\mathbb{m}}} = 0$. To represent these values where a human $\mathbb{h}$ observes a classroom for 15 minutes[9] with a model and where a single model $\mathbb{m}$ continues to observe for the remainder of the class (an additional 45 minutes), $\mathbf{K}_n = \{n_{o_{\mathbb{m}}} = 0, n_{o_{\mathbb{h}}} = 0, n_{o_{\mathbb{F}'}} = 1, n_{s_{\mathbb{m}}} = 6, n_{s_{\mathbb{h}}} = 0, n_{s_{\mathbb{F}'}} = 2, n_{r_{\mathbb{m}}} = 1, n_{r_{\mathbb{h}}} = 1, n_{o_{\mathbb{F}'}} = 0\}$ and where the variance components are solved similarly to the coefficients of Eq. 1.

6 Discussion

Figure 1 contains those MQI rubric items reported by Wang and Demszky, however all other items and evaluations can be found in the appendices. Contrasting panel (b) with panel (c) in Figure 1 reveals commonly used evaluation metrics obscure important aspects of model performance. However, as demonstrated in panels (c)-(f), there are methods that can be used to improve evaluation under label uncertainty. Many of these methods could be applied to annotated data prior to model training to improve data quality and support training (Gordon et al., 2022). The sentence-embedded Encoder models outperform humans on nearly every item in terms of inter-rater reliability metrics, improvements to ICC, and generalizable reliabilities as seen in standard metric tables of Figure 1 and in Appendix E.1. Using nearly any standardized combination of metrics across all items, Encoder models perform better than the *single highest performing expert human rater*. However, subsequent disattenuation analyses in Section 5.2 and 5.3 suggest that this Encoder model family's SOTA-level correlations on the EXPL item may have been spurious (likely identifying speech patterns associated with higher teacher performance, and not necessarily specific to explanations).

This is in stark contrast to the GPT models, which perform much worse than human raters. GPT models likely performed poorly in part due to the prompt length (Liu et al., 2023a) and the out-of-distribution nature of elementary school classroom discourse (McCoy et al., 2023). As GPT-style models increase in popularity, in use, and in sophistication, these methods can help identify sophistry and speciousness in third-party models even in the presence of low reliability. Like humans, models tended to choose a preferred rating value, and their deviations, conditionally informed by billions of parameters, are non-random. The GPT model families show a negative bias trend against Black teachers relative to White teachers on most items, as seen in the comparison of those models across panels (d) and (e) in Figure 1. Potentially more precisely, GPT models' rating centrality seemed to diminish when rating Black teachers, especially with the "reasoning" model, adding evidence that these foundation models may be sensitive to linguistic differences found in African-American English (AAE) (Hofmann et al., 2024b; Fleisig et al., 2024), possibly due to historical data or models' relative unfamiliarity with AAE (Rickford and King, 2016). These results alone should give pause to EdTech developers relying on prompt-engineering of foundation LLMs, as subtleties in biases exist in very complex tasks.

Being able to identify biases in cases of unreliable annotations is important, and researchers should resist the urge to ignore prompt engineered results from foundation models even if they fail to reject a null hypothesis. By performing more rigorous evaluations, researchers could crowdsource measuring model biases and behavior tendencies to help all users be more discerning of speciousness.

The Encoder models' designs, by contrast, were constructed to allow for multiple methods of interpretability and use by evaluating continuous windows of classroom discourse. This could be used for real-time diagnosis, interpretation, and supporting common understanding between teacher and coach. An example of such use can be found in Figure 9, where the continuous predictions for all models are displayed next to average human rating scores. Improvements to this process, combined with successful feature attribution, could boost validity and trust in model use for these high-stake scenarios.

As conducting actual human annotated classroom observation ratings is immensely expensive, the decision study analyses of Section 5.6 offer methods for estimating the improvement gained by using a model or model family. Parameterizing the decision conditions to reflect "human-in-the-loop" scenarios can even offer insight into whether the variation offered from automated ratings adds or detracts from human rating quality, offering a means of estimating research questions before more expensive trials.

Automated encoder LLMs could reduce the high costs of improving classroom observers' annotations and serve as a stepping stone to quality teacher development.[10] Education technologists and EdTech enthusiasts should be wary of foundation models' abilities to do out-of-distribution tasks, lest our "stochastic parrots" (Bender et al., 2021) start fires with their "embers of autoregression" (McCoy et al., 2023).

7 Limitations

The methods serve as a proof of concept for enhancing reliability in widespread and costly classroom evaluation tasks. Even though these models can

[9] For the MQI instrument, observation segments are 7.5 minutes long.

[10] Code for statistical models is available in the appendix and free for use.

perform better than a human given many accepted metrics, much more analysis and technological development is needed. Despite being best in class, these models should not be used in production in their current state. Even with a human in the loop, much more work must be done to ensure their readiness for possible assumed capabilities by end users. Far more important is that GPT style models are not used similarly, and this paper does not endorse their use for this or similar tasks.

By no means are the methods of this paper representative of psychometrics or even the best implementation of available methods. Rather it is intended to illustrate the potential for better quantifying behaviors in both labelers and models when we have uncertainty in labels. Psychometric models generally assume that the underlying latent variables are distributed normally across a population, which is usually a reasonable assumption with humans. But this assumption need not be true for models nor for all tasks. In this study, few models were estimated alongside humans to demonstrate how differently they behave under this assumption, but this paper provides no evidence that model abilities would be normally distributed for LLMs (e.g., latent constructs could follow multimodal distributions, depending on a family and pretraining, or follow a Normal-exponential-gamma distribution for shifts in metric-specific emergent behaviors). Were researchers interested in modeling learning in a larger population of models, other methods, such as, unipolar IRT models (Huang and Bolt, 2023), could potentially help for understanding between-model behaviors when the distribution. The usefulness of basic psychometric models presented is based on usefulness of the anthropomorphic comparisons we can reasonably make in the presence of uncertain labels.

The decision study of Section 5.6 does not have a within-observation-longitudinal parameterization and thus assumes that humans observing multiple segments of a class period do not necessarily need to observe the segments consecutively. While the MQI rubric is worded so as to be robust to within-lesson autocorrelation, actual lessons are obviously autocorrelated.

While many studies cited herein seek to generalize findings across all classrooms, we acknowledge that this cannot be done with the transcript data we use for this presented work, as it only consists of fourth and fifth-grade mathematics classrooms from the United States. Primarily, though the methods potentially possess broad applicability across all grades and subject areas, the current models have been exclusively developed and tested on transcripts from fourth and fifth-grade mathematics classrooms in U.S. public schools. Furthermore, the associated ratings and reliability metrics pertain solely to a subset of rating items on two specific rubrics, which may introduce limitations when addressing the more universal task of classroom evaluation.

These models were trained under the assumptions that the actual expert human ratings are not very reliable, that the alignment of the coordination of timing across rubrics and across transcripts is imperfect, that the discourse transcripts are imperfect, and that information is lost by keeping fixed sentence-level embeddings. While the methods outlined worked to extract a meaningful signal despite these challenges, it should be noted that the signal is still trained on noisy human ratings. If, on average, the raters had a particular bias, the model would carry that bias. For example, This is particularly true with the CLASS item ratings, as there were only 19 different raters used, compared to the 63 used for the MQI rubric items, and only had one rater per CLASS.

While not as severe as the GPT results, the encoder models did not avoid issues of racial bias. On the worst performing item for both human and encoder models, **MGEN**, all of the encoder models found spurious relationships in some language feature while overfitting with a negative bias against Black teachers. The reasons are likely to do with label sparcity: with so few examples of ratings in the higher categories in the training dataset, overfit on a potentially biased sample was not adequately controlled for, showing a microcosm of alignment to poor data that GPT exists in macrocosm. Fortunately for the encoders, many earlier data had already suggested that neither the models nor humans (see Appendix E.1 and (Hill et al., 2012b)) could sufficiently distinguish between the item's categories.

The encoder models could be improved through metalearning training, so they could be more adaptive to new instructional rubrics and classrooms. Transferrability is limited by the training regime and architecture just as much as lack of data for this task limits more robust model construction.

The encoder models removed transcription notes and intentionally did not use transcription information (such as identification of speaker) to best emulate what the functionality would be in a audio-input-only setup. While this is an authentic interpretation of the task, the transcription process was still done with humans. While direct input from audio would capture even more information (such as tone or long breaks in speaking for independent work), these models have not been trained to work with automated transcription.

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is Key: Disentangling Label Variation in Natural Language Processing

with Intra-Annotator Agreement. *arXiv preprint*. ArXiv:2301.10684 [cs].

Raymond J. Adams, Mark Wilson, and Wen-chung Wang. 1997. The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1):1–23. Place: US Publisher: Sage Publications.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2020. Sanity Checks for Saliency Maps. *arXiv preprint*. ArXiv:1810.03292 [cs, stat].

Elena Aguilar. 2013. Developing a Work Plan: How Do I Determine What to Do? In *The art of coaching: effective strategies for school transformation*, pages 119–144. Jossey-Bass, A Wiley Brand, San Francisco.

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 224–233, Seattle, Washington. Association for Computational Linguistics.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. Stop Measuring Calibration When Humans Disagree. *arXiv preprint*. ArXiv:2210.16133 [cs].

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting Predictive Probabilities: Model Confidence or Human Label Variation? *arXiv preprint*. ArXiv:2402.16102 [cs] version: 1.

Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2017. An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys.

Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73:101919.

Paul Bambrick-Santoyo. 2016. *Get better faster: a 90-day plan for coaching new teachers*. Jossey-Bass, A Wiley Brand, San Francisco, CA.

Paul Bambrick-Santoyo. 2018. *Leverage leadership 2.0: a practical guide to building exceptional schools*. Jossey-Bass, San Francisco, CA.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Isaac 1 Bejar, David M. Williamson, and and Robert J. Mislevy. 2006. Human Scoring. In *Automated Scoring of Complex Tasks in Computer-Based Testing*. Routledge. Num Pages: 34.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

David Blazar. 2018. Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence from Random Assignment of Teachers to Students. *Education Finance and Policy*, 13(3):281–309.

David Blazar, David Braslow, Charalambos Y. Charalambous, and Heather C. Hill. 2017. Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments. *Educational Assessment*, 22(2):71–94. Publisher: Routledge _eprint: https://doi.org/10.1080/10627197.2017.1309274.

Robert L. Brennan. 2001a. *Generalizability Theory*. Springer, New York, NY.

Robert L. Brennan. 2001b. Variability of Statistics in Generalizability Theory. In Robert L. Brennan, editor, *Generalizability Theory*, Statistics for Social Sciences and Public Policy, pages 179–213. Springer, New York, NY.

Robert L. Brennan. 2013. *Generalizability Theory*. Springer Science & Business Media. Google-Books-ID: nbHbBwAAQBAJ.

Derek C. Briggs and Mark Wilson. 2007. Generalizability in item response modeling. *Journal of Educational Measurement*, 44(2):131–155. Place: United Kingdom Publisher: Blackwell Publishing.

Jodi M. Casabianca. 2021. Digital Module 27: Hierarchical Rater Models. *Educational Measurement: Issues and Practice*, 40(4):103–104. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12478.

Jodi M. Casabianca, Daniel F. McCaffrey, Drew H. Gitomer, Courtney A. Bell, Bridget K. Hamre, and Robert C. Pianta. 2013. Effect of Observation Mode on Measures of Secondary Mathematics Teaching. *Educational and Psychological Measurement*, 73(5):757–783. Publisher: SAGE Publications Inc.

Eric P. Charles. 2005. The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets. *Psychological Methods*, 10(2):206–226. Place: US Publisher: American Psychological Association.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The Measure and Mismeasure of Fairness. *arXiv preprint*. ArXiv:1808.00023 [cs].

Chengyu Cui, Chun Wang, and Gongjun Xu. 2024. Variational Estimation for Multidimensional Generalized Partial Credit Model. *Psychometrika*.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv preprint*. ArXiv:2011.03395 [cs, stat].

Linda Darling-Hammond. 2014. What Can PISA Tell Us about U.S. Education Policy? *New England Journal of Public Policy*, 26(1).

Linda Darling-Hammond, Lisa Flook, Channa Cook-Harvey, Brigid Barron, and David Osher. 2020. Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2):97–140. Publisher: Routledge _eprint: https://doi.org/10.1080/10888691.2018.1537791.

Lawrence T. Decarlo. 2003. Using the PLUM procedure of SPSS to fit unequal variance and generalized signal detection models. *Behavior Research Methods, Instruments, & Computers*, 35(1):49–56.

Lawrence T. DeCarlo, YoungKoung Kim, and Matthew S. Johnson. 2011. A Hierarchical Rater Model for Constructed Responses, with a Signal Detection Rater Model. *Journal of Educational Measurement*, 48(3):333–356. Publisher: National Council on Measurement in Education.

Dorottya Demszky and Heather Hill. 2022. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. Publisher: arXiv Version Number: 1.

Dorottya Demszky and Heather Hill. 2023. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Dorottya Demszky and Jing Liu. 2023. M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, pages 59–69, New York, NY, USA. Association for Computing Machinery. Event-place: Copenhagen, Denmark.

Dorottya Demszky, Jing Liu, Heather C. Hill, Shyamoli Sanghi, and Ariel Chung. 2023. Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms. Technical report, Annenberg Institute at Brown University. Publication Title: EdWorkingPapers.com.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. Publisher: arXiv Version Number: 1.

Dorottya Demszky, Rose Wang, Sean Geraghty, and Carol Yu. 2024. Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, pages 632–644, New York, NY, USA. Association for Computing Machinery.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2022. Retiring Adult: New Datasets for Fair Machine Learning. *arXiv preprint*. ArXiv:2108.04884 [cs, stat].

Patrick J. Donnelly, Nathaniel Blanchard, Andrew M. Olney, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 218–227, New York, NY, USA. Association for Computing Machinery.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA. Association for Computing Machinery.

Thomas Eckes and Kuan-Yu Jin. Detecting Illusory Halo Effects in Rater- Mediated Assessment: A Mixture Rasch Facets Modeling Approach.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925,

Online. Association for Computational Linguistics.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. *arXiv preprint*. ArXiv:2406.08818 [cs] version: 1.

Shuai Gao. 2022. Système de traduction automatique neuronale français-mongol (historique, mise en place et évaluations) (French-Mongolian neural machine translation system (history, implementation, and evaluations) machine translation (hereafter abbreviated MT) is currently undergoing rapid development, during which less-resourced languages nevertheless seem to be less developed). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, pages 97–110, Avignon, France. ATALA.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint*. ArXiv:2209.14375.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–19, New York, NY, USA. Association for Computing Machinery.

Zaretta Hammond. 2015. *Culturally responsive teaching and the brain: promoting authentic engagement and rigor among culturally and linguistically diverse students*. Corwin, a SAGE company, Thousand Oaks, California. OCLC: ocn889185083.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv preprint*. ArXiv:1610.02413 [cs].

Mike Hardy. 2021. Toward Educator-focused Automated Scoring Systems for Reading and Writing. *arXiv preprint*. ArXiv:2112.11973 [cs].

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948. PMLR. ISSN: 2640-3498.

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. Publisher: Taylor & Francis, Ltd.

Heather C. Hill, Charalambos Y. Charalambous, David Blazar, Daniel McGinn, Matthew A. Kraft, Mary Beisiegel, Andrea Humez, Erica Litke, and Kathleen Lynch. 2012a. Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. *Educational Assessment*, 17(2-3):88–106. Publisher: Routledge _eprint: https://doi.org/10.1080/10627197.2012.715019.

Heather C. Hill, Charalambos Y. Charalambous, and Matthew A. Kraft. 2012b. When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, 41(2):56–64. Publisher: American Educational Research Association.

Andrew D. Ho and Thomas J. Kane. 2013. The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540957.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024a. AI generates covertly racist decisions about people based on their dialect. *Nature*, pages 1–8. Publisher: Nature Publishing Group.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024b. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint*. ArXiv:2403.00742 [cs].

Amin Hosseiny Marani, Joshua Levine, and Eric P.S. Baumer. 2022. One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 768–779, New York, NY, USA. Association for Computing Machinery.

Qi (Helen) Huang and Daniel M. Bolt. 2023. Unipolar IRT and the Author Recognition Test (ART). *Behavior Research Methods*.

Cassandra L. Jacobs, Ryan J. Hubbard, and Kara D. Federmeier. 2022. Masked language models directly encode linguistic uncertainty. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 225–228, online. Association for Computational Linguistics.

Xuejun (Ryan) Ji. 2023. *Using cross-classified mixed effects model for validation studies : a flexible and pragmatic validation method*. Ph.D. thesis, University of British Columbia.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, Roni Rabin, Jasmin Rubinovitz, Mac McAllister, Julia Wilkowski, David Choi, Roee Engelberg, Lidan Hackmon, Adva Levin, Rachel Griffin, Michael Sears, Filip Bar, Mia Mesar, Mana Jabbour, Arslan Chaudhry, James Cohan, Sridhar Thiagarajan, Nir Levine, Ben Brown, Dilan Gorur, Svetlana Grant, Rachel Hashimshoni, Jieru Hu, Dawn Chen, Kuba Dolecki, Canfer Akbulut, Maxwell Bileschi, Laura Culp, Wen-Xin Dong, Nahema Marchal, Kelsie Van Deman, Hema Bajaj Misra, Michael Duah, Moran Ambar, Avi Caciularu, Sandra Lefdal, Chris Summerfield, James An, Pierre-Alexandre Kamienny, Abhinit Mohdi, Theofilos Strinopoulous, Annie Hale, Wayne Anderson, Luis C Cobo, Niv Efron, Muktha Ananda, Shakir Mohamed, Maureen Heymans, Zoubin Ghahramani, Yossi Matias, Ben Gomes, and Lila Ibrahim. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach.

Thomas Kane, Heather Hill, and Douglas Staiger. 2015. National Center for Teacher Effectiveness Main Study: Version 4.

Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540959.

Thomas J. Kane and Douglas O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540960.

Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, Virtual Event Canada. ACM.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178.

Sean Kelly, Andrew M. Olney, Patrick Donnelly, Martin Nystrand, and Sidney K. D'Mello. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*, 47(7):451–464. Publisher: American Educational Research Association.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. *arXiv preprint*. ArXiv:2104.14337 [cs].

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *arXiv preprint*. ArXiv:1711.11279 [stat].

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv preprint*. ArXiv:1412.6980 [cs].

David Klahr. 2013. What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14075–14080. Publisher: Proceedings of the National Academy of Sciences.

Doug Lemov. 2021. *Teach like a champion 3.0: 63 techniques that put students on the path to college*, third edition edition. Jossey-Bass, a Wiley imprint, Hoboken, NJ.

Doug Lemov and Norman Atkins. 2015. *Teach like a champion 2.0: 62 techniques that put students on the path to college*, second edition edition. Jossey-Bass, San Francisco, CA.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv preprint*. ArXiv:2308.03281 [cs].

Peter Liljedahl, Tracy Johnston Zager, and Laura Wheeler. 2021. *Building thinking classrooms in mathematics: 14 teaching practices for enhancing learning: Grades K-12*. Corwin Mathematics. Corwin, Thousand Oaks, California London New Delhi Singapore.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the Middle: How Language Models Use Long Contexts. *arXiv preprint*. ArXiv:2307.03172 [cs].

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b.

G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint*. ArXiv:2303.16634 [cs].

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *arXiv preprint*. ArXiv:1705.07874 [cs, stat].

Panayota Mantzicopoulos, Brian F. French, and Helen Patrick. 2018. The Mathematical Quality of Instruction (MQI) in Kindergarten: An Evaluation of the Stability of the MQI Using Generalizability Theory. *Early Education and Development*, 29(6):893–908. Publisher: Routledge _eprint: https://doi.org/10.1080/10409289.2018.1477903.

Louis T. Mariano and Brian W. Junker. 2007. Covariates of the Rating Process in Hierarchical Models for Multiple Ratings of Test Items. *Journal of Educational and Behavioral Statistics*, 32(3):287–314.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve.

Samuel Messick. 1998. Test Validity: A Matter of Consequence. *Social Indicators Research*, 45(1/3):35–44. Publisher: Springer.

Eiji Muraki. 1992. A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16(2):159–176. Publisher: SAGE Publications Inc.

Daniel L. Murphy and S. Natasha Beretvas. 2015. A Comparison of Teacher Effectiveness Measures Calculated Using Three Multilevel Models for Raters Effects. *Applied Measurement in Education*, 28(3):219–236. Publisher: Routledge _eprint: https://doi.org/10.1080/08957347.2015.1042158.

Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. 2002. The Hierarchical Rater Model for Rated Test Items and Its Application to Large-Scale Educational Assessment Data. *Journal of Educational and Behavioral Statistics*, 27(4):341–384. Publisher: [American Educational Research Association, Sage Publications, Inc., American Statistical Association].

Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. 2008. *Classroom Assessment Scoring System (CLASS) Manual, K-3*. Paul H. Brookes Publishing Company. Google-Books-ID: NBeaGgAACAAJ.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. *arXiv preprint*. ArXiv:1709.02012 [cs, stat].

Martyn Plummer. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Working Papers*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

John R. Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4):948–988.

Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv preprint*. ArXiv:1811.10154 [cs, stat].

Borhan Samei, Andrew M. Olney, Sean Kelly, Martin Nystrand, Sidney D'Mello, Nathan Blanchard, Xiaoyi Sun, Marcy Glaus, and Art Graesser. 2014. Domain Independent Assessment of Dialogic Properties of Classroom Discourse. Technical report. Publication Title: Grantee Submission ERIC Number: ED566380.

Jon Saphier, Mary Ann Haley-Speca, and Robert Gower. 2008. *The skillful teacher: building your teaching skills*, 6th ed edition. Research for Better Teaching, Acton, Mass.

Daniel L. Schwartz, Jessica M. Tsang, and Kristen P. Blair. 2016. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*, first edition edition. Norton books in education. W.W. Norton & Company, New York.

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

Robert E. Slavin. 2002. Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, 31(7):15–21. Publisher: American Educational Research Association.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2020. Learning Controllable Fair Representations. *arXiv preprint*. ArXiv:1812.04218 [cs, stat].

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR. ISSN: 2640-3498.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

R Core Team. R: A Language and Environment for Statistical Computing.

Jiarui Wang, Richong Zhang, Junfan Chen, Jaein Kim, and Yongyi Mao. 2022. Text style transferring via adversarial masking and styled filling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7654–7663, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Albert Webson, Alyssa Loo, Qinan Yu, and Ellie Pavlick. 2023. Are Language Models Worse than Humans at Following Prompts? It's Complicated. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7662–7686, Singapore. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do Prompt-Based Models Really Understand the Meaning of Their Prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated Evaluation of Classroom Instructional Support with LLMs and BoWs: Connecting Global Predictions to Specific Feedback. *arXiv preprint*. ArXiv:2310.01132 [cs].

Grover J. Whitehurst, Matthew M. Chingos, and Katharine M. Lindquist. 2014. Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts. Technical report, Brookings Institution. Publication Title: Brookings Institution ERIC Number: ED553815.

Stefanie A. Wind. 2019. Nonparametric Evidence of Validity, Reliability, and Fairness for Rater-Mediated Assessments: An Illustration Using Mokken Scale Analysis. *Journal of Educational Measurement*, 56(3):478–504. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jedm.12222.

Stefanie A. Wind and Wenjing Guo. 2019. Exploring the Combined Effects of Rater Misfit and Differential Rater Functioning in Performance Assessments. *Educational and Psychological Measurement*, 79(5):962–987. Publisher: SAGE Publications Inc.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333. PMLR. ISSN: 1938-7228.

Shengjia Zhao and Stefano Ermon. 2021. Right Decisions from Wrong Predictions: A Mechanism Design Alternative to Individual Calibration. *arXiv preprint*. ArXiv:2011.07476 [cs, math, stat].

A Observation Instrument Item Descriptions and Distributions

For each of the observation instruments, the abbreviation codes used in this study are listed with the expanded names in Table 1. The distributions of scores across all items for all rater families are in Figure 3. The CLASS rubric has 12 items on a scale from 1 to 7, rated at 15 minute intervals. The MQI rubric has 13 items on a scale from 1 to 3, rated at 7.5 minute intervals.

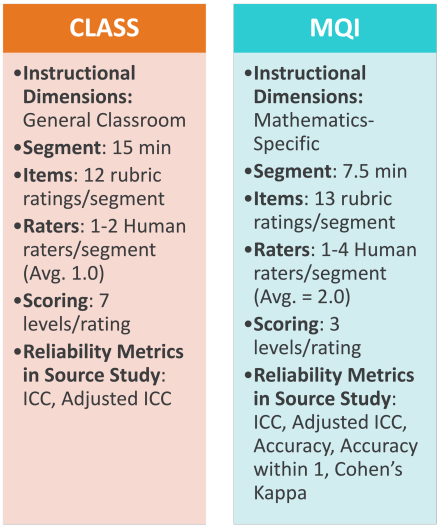

Figure 2: Overview of technical details the two instructional frameworks used for evaluating instruction.

Item Description	**Abbreviation**
CLASS Instrument Items	
Classroom Positive Climate	CLPC
Classroom Negative Climate	CLNC
Teacher Sensitivity	CLTS
Regard for Student Perspective	CLRSP
Behavior Management	CLBM
Productivity	CLPRDT
Instructional Learning Formats	CLILF
Content Understanding	CLCU
Applied Problem Solving	CLAPS
Quality of Feedback	CLQF
Instructional Dialogue	CLINSTD
Student Engagement	CLSTENG
MQI Instrument Items	
Enacted Task Cognitive Activation	ETCA
Explanations	EXPL
Imprecision in Language or Notation	LANGIMP
Lack of Clarity in Presentation of Mathematical Content	LCP
Linking and Connections	LINK
Major Mathematical Errors	MAJERR
Developing Mathematical Generalizations	MGEN
Mathematical Language	MLANG
Multiple Procedures or Solution Methods	MMETH
Remediation of Student Errors and Difficulties	REMED
Student Mathematical Questioning and Reasoning	SMQR
Students Provide Explanations	STEXPL
Responding to Student Mathematical Productions	USEPROD

Table 1: CLASS and MQI item descriptions and corresponding abbreviations.

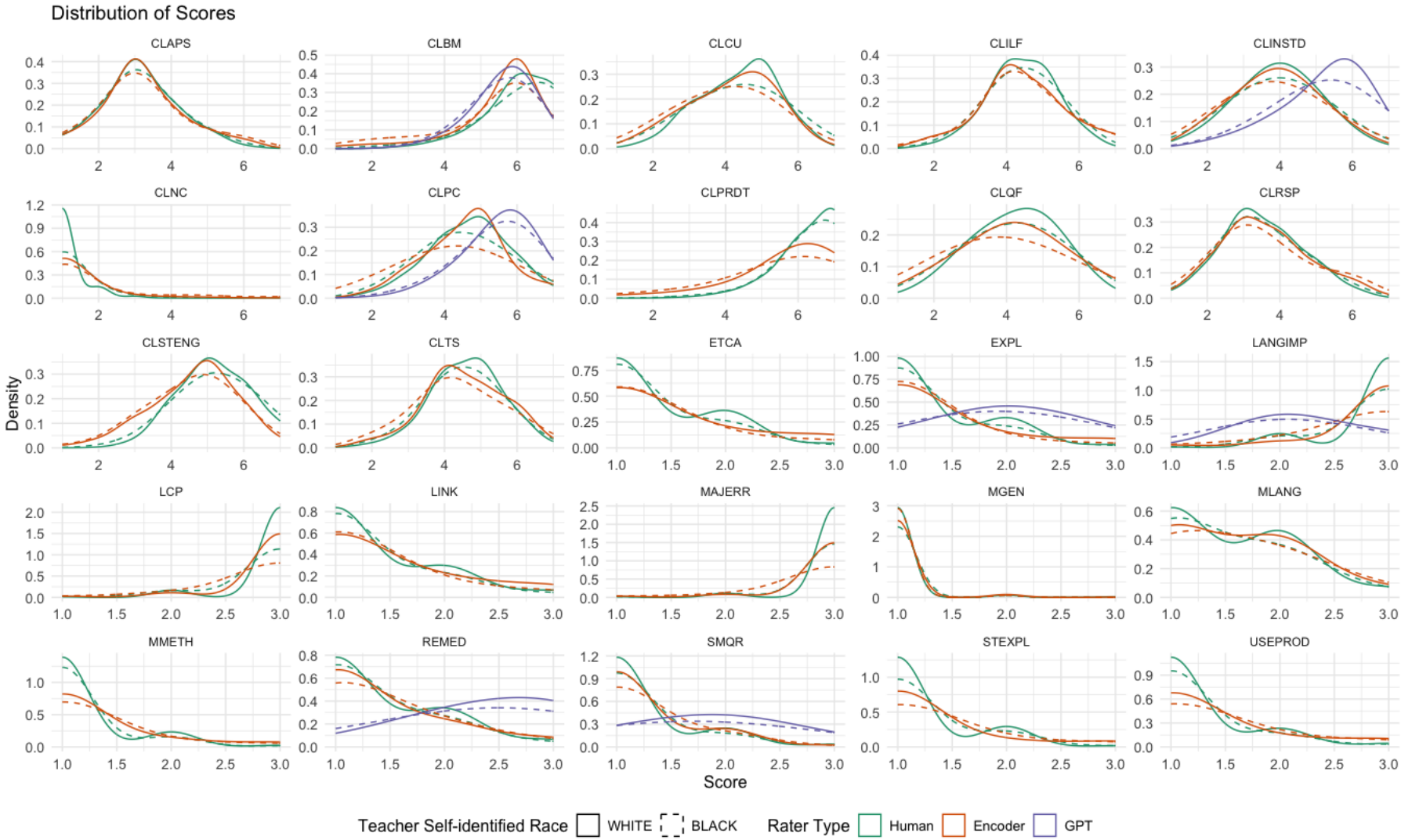

Figure 3: Distribution of rater scores for each of the 25 instrument items for all rater families.

B MQI Instrument

B.1 MQI Instrument Properties

For our purposes, the MQI instrument has a few unique properties that warrant further analysis, as the instrument may have some qualitative attributes that may influence human raters.

The MQI ratings are written to identify the presence of a behavior and then, if present, report the magnitude or quality of its presence, doing so repeatedly at regular intervals throughout the lesson (in this case, 7.5 minutes). This shortened window with simpler targets provides an opportunity for training a model for real-time use (rather than an arbitrary interval) to find different features across a single lesson, as shown in Figure 9.

The version of the MQI for which data is available in the NCTE dataset is ternary, in contrast to the current MQI version, which is quaternary. The lowest rating on the ternary MQI scale is a combination of the lowest two ratings on the quaternary, meaning the present data cannot distinguish between whether the attribute described in each item is "Not present" or "Low". [11] This ternary classification scheme creates non-normal distributions as seen in Figure **??**, which will need to inform models and methods during quantitative analysis.

This is unfortunate because the difference between these two categories are "None." And "Brief content error, instance of imprecision, lack of clarity. Does not obscure the mathematics of the segment," respectively (for the Errors and Imprecision domain in Hill et al. and second MQI-only factor in Blazar et al.: **MAJERR**, **LANGIMP**, **LCP**).

B.2 Possible Effects of Negative-worded Items

The MQI is unique in having a separate domain of items that try to capture aspects of *poor* mathematical instruction. Unlike most items in observation rubrics, the MQI has three items that are worded in the negative direction, specifically, higher scores on the **MAJERR**, **LANGIMP**, and **LCP** items indicate worse performance.[12] It is possible that looking for negative attributes may make these items more susceptible to different rater biases. A partial description of the potential impact of this rubric attribute for the LCP item found in Appendix B.2 with further details.

Of note, the LCP item is particularly subjective. In the documentation and training provided for the MQI, You have to ask: "What, mathematically, was the teacher trying to say?" This is already problematic, as it is asking for observers to use their judgment to determine what the teacher was "trying to say." The subjectivity increases further for observers who may not be as familiar with African-American Vernacular English (AAVE). The subjectivity is further mixes lack of content clarity (lack of clarity explaining math) with lack of directional clarity (unclear instructions for an activity, which is typically associated with items addressing classroom management), as stated in the MQI rubric:

> *Teacher's launch of a task/activity lacks clarity (the "launch" is the teacher's effort to get the mathematical tasks/activities into play). If the launch is problematic, score for the launch plus amount of time students are confused/off-task/engaging in non-productive explorations... [Example:] Garbling a task launch, e.g., by asking initially "How much TV is watched in the US?" when students really must draw a graph to show "How many TVs in US vs. Europe vs. rest of the world?*

Instructing observers to score based on the "amount of time students are confused/off-task/engaging in non-productive explorations", is more likely to capture problems with classroom management and directional lack of clarity, not mathematical lack of clarity, compounded by the request for raters to guess what the teachers were trying to say and training instructions that let raters "code Lack of Clarity even with correction". This mix of observational cues and overlapping constructs makes this item particularly susceptible to individual rater biases.[13]

Indeed, while not reported in this paper explicitly, we identified that one rater in particular rated Black teachers much more harshly on these, especially on LCP, providing some evidence that some items can be more prone to rater biases, even with research-quality observers and calibration.

B.3 Prior work on Rater Fairness with MQI

Recent work has begun to look at rater biases, including racial bias, in these data and with the MQI instrument. Ji (2023) uses cross-classified mixed effects models for analysis and evaluation, which seeks to answer similar questions through combining G-theory and IRT estimations (Briggs and Wilson, 2007). However, the helpfulness of this study is

[11] There is one exception, which the original authors of the Appendix adjusted for: the **USEPROD** item is replaced by the **MATCON** item, with the correction of combining the lowest two categories.

[12] In the analyses of this paper, these will be reverse coded, as will the one negative CLASS item **CLNC**

[13] As a note, the skill of providing clear directions, foundational to establishing a well-managed classroom, is also not included the CLASS instrument's "Behavior Management" item, suggesting that neither of these instruments is perfectly designed to address root causes of instructional shortcomings and thus may be inadequate as tools for coaching and developing skills in teachers.

limited by data selection decisions: it eliminates 23% of MQI items (all of the second MQI factor in (Blazar et al., 2017)) without explanation; it only uses 21% of available classroom observations (from a single year) and by so doing also eliminates 43% of the study's raters; it then truncates the class lengths to 45 minutes thus removing another 20% of the remaining data observations, and when evaluating for differences in teacher race, combines all non-white races/ethnicities into a single category, removing meaningful inference from the contrast. These decisions to use only 13% of available data would lead to a model with better fit, as all of those removals simplify trends in the data, indirectly suggesting that the mixed effects model constructions used are not robust to the complete set of observations (Murphy and Beretvas, 2015) and are therefore inadequate for our purposes here.

C Encoder Family Construction

Pretraining and training/fine-tuning regimes can have significant effects on model performance (D'Amour et al., 2020), so our family of models sought to exploit this by using three different pretrainings for sentence-level embeddings and including variations on training regimes (e.g., different checkpoints), the summary of these variations can be found in Table 2. Thus, the encoder family of models designed for this study share the same architecture,[14] training and held-out test sets, differing only as outlined in Table 2.

[Another forthcoming paper to be under review] explores this protocol in greater depth, showing that the extreme training and treatment of data noise can achieve SOTA and "super-human" results on a variety of sentence embedding pretrainings, with a more complete set of training

All results were run on a completely held out test set of entire classroom transcripts. No analyses were conducted using the held-out test set until after all models in the model family were trained, thus preserving the integrity of the study.

C.0.1 Encoder Model Preprocessing

As mentioned in Section 4, preprocessing of the transcript data was intentionally minimal, replacing bracketed transcription notes (e.g. `[cross-talk]`) with `[inaudible]`. For this study, the transcript was not annotated denote whether a teacher or a student is speaking to reflect the broadest future use case of general classroom microphones. In other words, this family of models does not know who is speaking, and the results of this decision are evident in the models' relative underperformance in two MQI items that distinguish between teacher explanations (EXPL) and student explanations (STEXPL), a trend that might be evident in the validity demonstration in Figure 9, where models may be responding nearly identically to/failing to distinguish between these two items.

To align transcripted class segments to human observation ratings, transcripts were equipartitioned at the word-level across the maximum number of lesson segments for which there were human annotations available, and estimated timestamps were made across sentences by linear interpolation weighted by word count.

C.1 Sentence-level Embeddings

One key difference to other studies using these same transcripts is the choice to parse the utterances at the sentence level. Sentences, rather than individual words or long, uninterrupted utterances, are the key unit of meaning for interpretability of models for classroom discourse. The downstream tasks are a key decision for this choice. Sentence level parsing anticipates meaningful feature attribution studies (Sundararajan et al., 2017) to further investigate construct validity.

Parsing at the sentence level both augments the total number of unique observations in the data and, by creating more standardization in sequence lengths prior to sentence-embedding, the variation in the density of semantic information is reduced.

The model takes as input an approximate 12 min rolling window of class text (stepping at each sentence), and simultaneously predicts ratings for each of the 12 CLASS dimensions, 13 of the MQI dimensions for rounded-rolling average scores for that time window. Each model is multi-task predicting all 25 scores simultaneously for each of the MQI and CLASS items. This multi-task training takes advantage of the interrelated skills of teaching that may be implicit in human ratings. Over one million unique observations from fewer than 1,600 unique classroom transcripts were generated, with rolling windows representing each observation. Training-val-test splits of this data were 75/15/10, stratified at the classroom level.

Classroom transcripts are extremely long, with thousands of sentences, and with classes having tokens in the hundreds of thousands. Sentence-level inputs could capture the relationship between something a teacher says and something a student says five minutes later without incurring large costs associated with sequence length. These long-range dependencies are needed to identify some of the instructional constructs being measured.

Raw class transcripts also have a lot of noise: content that is unrelated to any of the tasks, including fillers, self-corrections, interruptions and self-interruptions, sentences that are partially repeated or emphasized, text that requires being able to refer to a visual cue in the classroom, etc. While sentence level embeddings lose information relative

[14] One model, "un2", has a slightly different architecture, differing in the number of attention heads.

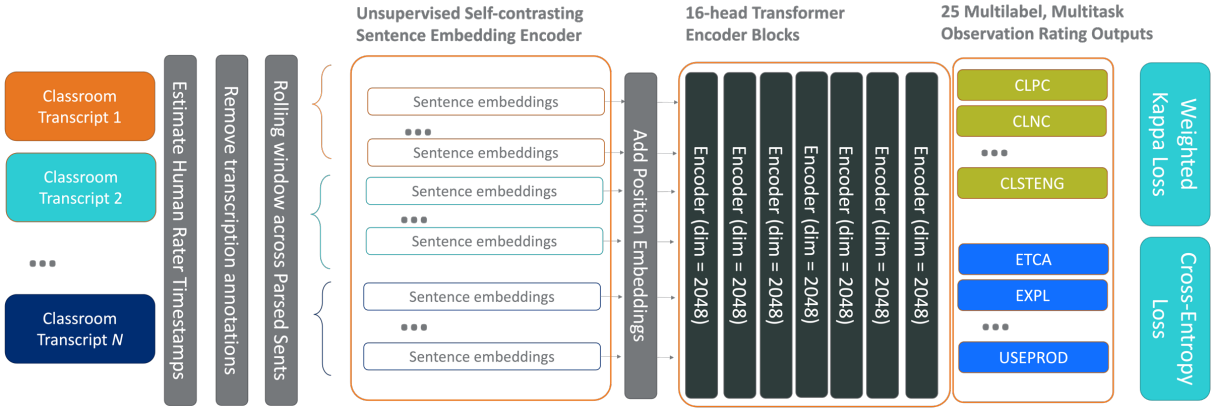

Figure 4: Model Pipeline: General sentence-encoder model architecture.

Model	Pretrained Embedding	Layer	Attn. Heads	Train Epochs	Dropout
un1	Unsupervised SimCSE (Gao, 2022)		32	3	75
un2	Unsupervised SimCSE (Gao, 2022)		16	4	75
un3	Unsupervised SimCSE (Gao, 2022)		32	8	75
e5	E5 (Wang et al., 2022)		32	2	15
gte	GTE (Li et al., 2023)		32	4	65

Table 2: Encoder Within-family differences: Summary of basic differences within the Encoder family of models. Detailed information about training and architecture can be found in Appendix C.3.

to subword tokenizations, this loss of information may mitigate disproportionate effects of idiosyncratic speaking styles.

C.1.1 Embedding Model Selection

To save on compute, static embeddings were precomputed. To represent the very noisy transcript data, we have to be careful in using sentence-embeddings, as they decrease the completeness of the information captured. We tested sentence-level embeddings using across different pretrained embedding models accessed through Huggingface on a subset of the training data for a small random selection of target measures:

- `unsup-simcse-roberta-large`: from princeton-nlp (Gao, 2022), was pretrained using unsupervised contrastive sentence representations. simCSE
- `sup-simcse-roberta-large`: from princeton-nlp (Gao, 2022), was pretrained using supervised training. At the writing of this paper, we did not yet have a converged model with reportable results. simCSE
- `e5-large-v2`: from intfloat (Wang et al., 2022), pretrained using weakly supervised contrastive sentence representations with sentence pair training. e5-large-v2
- `gte-large`: from thenlper (Li et al., 2023), pretrained using multistage contrastive sentence representations. gte-large

The first three models had significantly reduced performance, compared to our sentence embedding model of choice, SimCSE (Gao, 2022), which uses unsupervised self-contrasting learning to improve sentence-level representations of words.

C.2 Model Architecture

C.3 Encoder Model Training and Description

Models were built and trained in pytorch,[15] largely based on the Encoder modules available. Each model was trained on a single L4 GPU in Google Colab. Each epoch took about 4.25 hours:

- 8 transformer encoder layers
- 25 total classifier heads (with a single dense layer each) for each task (using double objective functions, results 50 total loss calculations backpropagated.)
- All encoder layer parameters are shared by objectives, but the trainable parameters of the single dense layer classification heads are specific to each item.
- Attention heads: 32. Since a lot of semantic information were needed to be extracted from

[15] https://pytorch.org/docs/stable/generated/torch.nn.TransformerEncoder.html

GPT Model	Name	Prompt Info	Output
N	Numeric	Item Overview	Single Number
ND	Numeric w/ Description	Rubric Descriptions of Score Categories	Single Number
NR	Numeric after Reasoning	Item Overview and CoT instructions	Reasoning and Number

Table 3: GPT Within-family model differences: Details for the GPT/Decoder models can be found in the original paper (Wang and Demszky, 2023).

within each embedding and its neighbors, supporting an increase in multi-head self-attention mechanisms.

- Hidden dimension: 2048

C.3.1 Preventing Overfit within the Model

An abnormally high 0.75 Dropout rate was the primary regularization technique to avoid overfit in a noisy dataset with non-gold labels.

- **Optimizer: Adamax**: defined in the original paper by Kingma and Ba (2017), this is a variant of Adam that replaces the L2 norm of the gradients with the L-infinity norm which provides stability in sparse gradients resulting from the droupout. Additionally, its initial momentum and second derivative momentum are limited slightly to 0.78 and 0.9, respectively, to prevent overfitting, but increasing training time, and increased the weight decay to 0.0003 similarly.
- **Learning Rate**: initial learning rate was set to 2.5e-5
- **Gradient clipping**: set to 4 (instead of the typical 1), since we did not want an unusual batch to , but recognizing that we need to capture as much info as we can from our optimizer
- **Learning rate schedule**: Using chaining, began linear from zero with warmup, a 1,000 step linear ramp, followed by exponential decay with gamma = 0.9995) with CosineAnnealingWarmRestarts scheduling with annealing cycles cutting frequency by a third each time. We have initial data to suggest that using a cyclic learning rate improves model performance.

D GPT Model Family

D.1 Model construction

Detailed descriptions of the three models and data generated by them can be found in the original paper and accompanying websites Wang and Demszky[16] which examples for how the three models differ. A brief summary of those differences can be found in Table 3.

D.1.1 GPT Model Preprocessing

In contrast to the Encoder model preprocessing, a preliminary analysis was conducted by Wang and Demszky to identify the highest quality 7.5-minute segments available in the dataset, as defined by fewest transcriber notes. The models are provided the discrourse from these selections and also information about the subset of items they provide ratings for, including four items from the MQI (**EXPL**, **LANGIMP**, **REMED**, **SMQR**).

E Reliability Metrics

ICC calculations were reproduced using the following multilevel model where lesson l scores for each rubric item are nested within teachers k:

$$ITEM_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}, \tag{11}$$

and then calculate the ICC and Adjusted ICC

$$ICC = \frac{\text{var}\left(\mu_k\right)}{\text{var}\left(\mu_k\right) + \frac{\text{var}(\varepsilon_{lk})}{n_l}}, \tag{12}$$

where $n_l = 1$ for ICC and where $n_l = 6$ for Adjusted ICC following the original study. Full results of human baselines and comparisons against the various models can be found in Appendix E.1.

E.1 Full Results

Table 4 contains the full results calculations referenced in Section 5.1. The metric symbols found in the table are as follows: **CK**: Cohen's κ; **QWK**: Quadratic Weighted Kappa; **ACC**: percent exact agreement; **ACC2**: percent agreement within 1 category; **ICC** and **AdjICC**: intraclass correlation and adjusted intraclass correlation (with nested staging in Eq. 12; **CORR**: ρ, Pearson's correlation; **CORp**: $\mathbf{r_s}$, Spearman's rank correlation, **CORt**: τ, Kendall's rank correlation.

[16]The automated rating data was retrieved from https://github.com/rosewang2008/zero-shot-teacher-feedback/tree/main

Table 4: Full Agreement Metrics

Instrument	Item	Metric	Human	Encoder	un1	un2	un3	gte	e5	GPT	N	ND	NR
MQI	LINK	CK	0.31	0.36	0.38	0.36	0.42	0.36	0.36	NA	NA	NA	NA
MQI	LINK	QWK	0.41	0.55	0.60	0.59	0.60	0.54	0.54	NA	NA	NA	NA
MQI	LINK	ACC	0.70	0.71	0.73	0.73	0.75	0.70	0.70	NA	NA	NA	NA
MQI	LINK	ACC2	0.97	0.97	0.97	0.98	0.98	0.98	0.98	NA	NA	NA	NA
MQI	LINK	CORR	0.41	0.56	0.61	0.60	0.61	0.54	0.54	NA	NA	NA	NA
MQI	LINK	ICC	0.14	0.13	0.13	0.14	0.13	0.14	0.14	NA	NA	NA	NA
MQI	LINK	AdjICC	0.49	0.48	0.48	0.50	0.48	0.49	0.49	NA	NA	NA	NA
MQI	EXPL	CK	0.23	0.27	0.30	0.33	0.26	0.28	0.28	0.05	0.02	0.08	0.04
MQI	EXPL	QWK	0.27	0.44	0.49	0.48	0.47	0.44	0.44	0.04	0.04	0.12	-0.02
MQI	EXPL	ACC	0.70	0.71	0.73	0.72	0.72	0.73	0.73	0.32	0.31	0.42	0.18
MQI	EXPL	ACC2	0.98	0.97	0.97	0.98	0.97	0.97	0.97	0.86	0.95	0.92	0.67
MQI	EXPL	CORR	0.27	0.45	0.51	0.49	0.49	0.45	0.45	0.07	0.08	0.18	-0.06
MQI	EXPL	ICC	0.15	0.16	0.14	0.13	0.14	0.14	0.14	0.16	0.13	0.13	0.12
MQI	EXPL	AdjICC	0.52	0.54	0.49	0.48	0.49	0.49	0.49	0.53	0.47	0.47	0.45
MQI	MMETH	CK	0.41	0.27	0.46	0.37	0.33	0.25	0.25	NA	NA	NA	NA
MQI	MMETH	QWK	0.47	0.44	0.46	0.53	0.54	0.45	0.45	NA	NA	NA	NA
MQI	MMETH	ACC	0.85	0.79	0.88	0.85	0.84	0.77	0.77	NA	NA	NA	NA
MQI	MMETH	ACC2	0.99	0.97	0.99	0.98	0.98	0.97	0.97	NA	NA	NA	NA
MQI	MMETH	CORR	0.47	0.48	0.48	0.57	0.57	0.51	0.51	NA	NA	NA	NA
MQI	MMETH	ICC	0.15	0.18	0.14	0.15	0.15	0.16	0.16	NA	NA	NA	NA
MQI	MMETH	AdjICC	0.51	0.56	0.49	0.52	0.52	0.53	0.53	NA	NA	NA	NA
MQI	MGEN	CK	0.15	0.18	0.28	0.38	0.25	0.16	0.16	NA	NA	NA	NA
MQI	MGEN	QWK	0.20	0.26	0.33	0.53	0.30	0.22	0.22	NA	NA	NA	NA
MQI	MGEN	ACC	0.95	0.94	0.96	0.97	0.96	0.94	0.94	NA	NA	NA	NA
MQI	MGEN	ACC2	0.99	0.99	1.00	1.00	1.00	0.99	0.99	NA	NA	NA	NA
MQI	MGEN	CORR	0.20	0.26	0.36	0.54	0.34	0.22	0.22	NA	NA	NA	NA
MQI	MGEN	ICC	0.04	0.03	0.01	0.02	0.01	0.02	0.02	NA	NA	NA	NA
MQI	MGEN	AdjICC	0.18	0.18	0.06	0.12	0.07	0.11	0.11	NA	NA	NA	NA
MQI	MLANG	CK	0.23	0.37	0.40	0.40	0.45	0.33	0.33	NA	NA	NA	NA
MQI	MLANG	QWK	0.33	0.48	0.49	0.52	0.53	0.46	0.46	NA	NA	NA	NA
MQI	MLANG	ACC	0.59	0.65	0.68	0.68	0.71	0.62	0.62	NA	NA	NA	NA
MQI	MLANG	ACC2	0.98	0.99	0.99	0.98	0.99	0.98	0.98	NA	NA	NA	NA
MQI	MLANG	CORR	0.33	0.48	0.50	0.52	0.53	0.48	0.48	NA	NA	NA	NA
MQI	MLANG	ICC	0.07	0.08	0.06	0.05	0.06	0.07	0.07	NA	NA	NA	NA
MQI	MLANG	AdjICC	0.31	0.34	0.27	0.25	0.26	0.29	0.29	NA	NA	NA	NA
MQI	REMED	CK	0.26	0.28	0.27	0.32	0.38	0.24	0.24	0.00	-0.01	0.02	0.00
MQI	REMED	QWK	0.32	0.41	0.45	0.48	0.42	0.40	0.40	0.00	0.00	0.01	0.00
MQI	REMED	ACC	0.66	0.68	0.67	0.67	0.75	0.66	0.66	0.15	0.09	0.28	0.07
MQI	REMED	ACC2	0.96	0.97	0.94	0.95	0.98	0.97	0.97	0.58	0.51	0.77	0.43
MQI	REMED	CORR	0.32	0.41	0.47	0.49	0.45	0.40	0.40	-0.01	-0.02	0.01	-0.02
MQI	REMED	ICC	0.14	0.15	0.13	0.13	0.13	0.14	0.14	0.13	0.11	0.12	0.11
MQI	REMED	AdjICC	0.49	0.52	0.47	0.48	0.46	0.49	0.49	0.48	0.43	0.45	0.43
MQI	USEPROD	CK	0.24	0.29	0.29	0.30	0.29	0.32	0.32	NA	NA	NA	NA
MQI	USEPROD	QWK	0.33	0.46	0.48	0.47	0.49	0.47	0.47	NA	NA	NA	NA
MQI	USEPROD	ACC	0.76	0.74	0.74	0.79	0.75	0.74	0.74	NA	NA	NA	NA
MQI	USEPROD	ACC2	0.98	0.96	0.95	0.97	0.95	0.96	0.96	NA	NA	NA	NA
MQI	USEPROD	CORR	0.33	0.47	0.51	0.48	0.51	0.49	0.49	NA	NA	NA	NA
MQI	USEPROD	ICC	0.23	0.22	0.20	0.21	0.20	0.20	0.20	NA	NA	NA	NA
MQI	USEPROD	AdjICC	0.64	0.62	0.60	0.61	0.59	0.61	0.61	NA	NA	NA	NA
MQI	MAJERR	CK	0.23	0.21	0.31	0.27	0.27	0.19	0.19	NA	NA	NA	NA
MQI	MAJERR	QWK	0.27	0.36	0.35	0.50	0.50	0.29	0.29	NA	NA	NA	NA
MQI	MAJERR	ACC	0.91	0.89	0.93	0.90	0.92	0.87	0.87	NA	NA	NA	NA
MQI	MAJERR	ACC2	0.99	0.99	1.00	0.99	0.99	0.98	0.98	NA	NA	NA	NA
MQI	MAJERR	CORR	0.27	0.37	0.37	0.50	0.51	0.32	0.32	NA	NA	NA	NA
MQI	MAJERR	ICC	0.09	0.08	0.07	0.07	0.07	0.07	0.07	NA	NA	NA	NA
MQI	MAJERR	AdjICC	0.37	0.34	0.31	0.30	0.30	0.30	0.30	NA	NA	NA	NA
MQI	LANGIMP	CK	0.25	0.20	0.31	0.24	0.22	0.17	0.17	0.00	0.02	-0.03	0.00
MQI	LANGIMP	QWK	0.29	0.34	0.35	0.43	0.41	0.33	0.33	0.00	0.03	-0.03	0.02
MQI	LANGIMP	ACC	0.80	0.79	0.85	0.82	0.82	0.75	0.75	0.31	0.26	0.38	0.29
MQI	LANGIMP	ACC2	0.99	0.98	0.99	0.98	0.98	0.97	0.97	0.98	0.97	0.98	0.99
MQI	LANGIMP	CORR	0.29	0.34	0.39	0.44	0.42	0.35	0.35	0.01	0.07	-0.06	0.03
MQI	LANGIMP	ICC	0.12	0.12	0.11	0.10	0.11	0.12	0.12	0.12	0.11	0.11	0.12
MQI	LANGIMP	AdjICC	0.45	0.46	0.43	0.41	0.43	0.45	0.45	0.46	0.43	0.42	0.44
MQI	LCP	CK	0.18	0.18	0.26	0.33	0.17	0.14	0.14	NA	NA	NA	NA
MQI	LCP	QWK	0.23	0.31	0.32	0.53	0.38	0.26	0.26	NA	NA	NA	NA
MQI	LCP	ACC	0.86	0.84	0.88	0.89	0.88	0.82	0.82	NA	NA	NA	NA

Table 4: Full Agreement Metrics *(continued)*

Instrument	Item	Metric	Human	Encoder	un1	un2	un3	gte	e5	GPT	N	ND	NR
MQI	LCP	ACC2	0.99	0.98	0.99	0.98	0.99	0.98	0.98	NA	NA	NA	NA
MQI	LCP	CORR	0.23	0.32	0.36	0.54	0.38	0.26	0.26	NA	NA	NA	NA
MQI	LCP	ICC	0.14	0.14	0.08	0.10	0.08	0.08	0.08	NA	NA	NA	NA
MQI	LCP	AdjICC	0.49	0.49	0.34	0.39	0.35	0.35	0.35	NA	NA	NA	NA
MQI	STEXPL	CK	0.36	0.27	0.25	0.30	0.24	0.32	0.32	NA	NA	NA	NA
MQI	STEXPL	QWK	0.39	0.41	0.42	0.48	0.45	0.42	0.42	NA	NA	NA	NA
MQI	STEXPL	ACC	0.80	0.77	0.77	0.80	0.78	0.78	0.78	NA	NA	NA	NA
MQI	STEXPL	ACC2	0.99	0.97	0.96	0.98	0.97	0.97	0.97	NA	NA	NA	NA
MQI	STEXPL	CORR	0.39	0.44	0.45	0.49	0.48	0.45	0.45	NA	NA	NA	NA
MQI	STEXPL	ICC	0.29	0.28	0.25	0.24	0.24	0.25	0.25	NA	NA	NA	NA
MQI	STEXPL	AdjICC	0.71	0.70	0.67	0.66	0.65	0.66	0.66	NA	NA	NA	NA
MQI	SMQR	CK	0.24	0.26	0.23	0.31	0.30	0.28	0.28	0.04	0.07	0.07	-0.01
MQI	SMQR	QWK	0.30	0.36	0.42	0.42	0.38	0.35	0.35	0.08	0.08	0.09	0.09
MQI	SMQR	ACC	0.76	0.75	0.75	0.79	0.76	0.74	0.74	0.39	0.41	0.47	0.24
MQI	SMQR	ACC2	0.98	0.99	0.96	0.97	0.99	0.99	0.99	0.91	0.91	0.90	0.94
MQI	SMQR	CORR	0.30	0.36	0.43	0.43	0.39	0.35	0.35	0.14	0.12	0.12	0.23
MQI	SMQR	ICC	0.18	0.20	0.18	0.18	0.18	0.18	0.18	0.20	0.18	0.17	0.17
MQI	SMQR	AdjICC	0.57	0.59	0.56	0.57	0.56	0.57	0.57	0.59	0.57	0.56	0.55
MQI	ETCA	CK	0.23	0.30	0.29	0.35	0.29	0.31	0.31	NA	NA	NA	NA
MQI	ETCA	QWK	0.31	0.49	0.50	0.53	0.50	0.49	0.49	NA	NA	NA	NA
MQI	ETCA	ACC	0.67	0.68	0.68	0.72	0.67	0.70	0.70	NA	NA	NA	NA
MQI	ETCA	ACC2	0.98	0.97	0.96	0.98	0.95	0.98	0.98	NA	NA	NA	NA
MQI	ETCA	CORR	0.31	0.50	0.52	0.53	0.54	0.49	0.49	NA	NA	NA	NA
MQI	ETCA	ICC	0.20	0.21	0.19	0.19	0.19	0.19	0.19	NA	NA	NA	NA
MQI	ETCA	AdjICC	0.60	0.62	0.59	0.58	0.59	0.58	0.58	NA	NA	NA	NA
CLASS	CLPC	CK	1.00	0.26	0.31	0.32	0.23	0.30	0.30	0.01	-0.01	0.03	-0.01
CLASS	CLPC	QWK	1.00	0.62	0.72	0.65	0.63	0.65	0.65	-0.02	-0.02	0.02	-0.03
CLASS	CLPC	ACC	1.00	0.45	0.47	0.49	0.45	0.50	0.50	0.23	0.21	0.27	0.20
CLASS	CLPC	ACC2	1.00	0.88	0.94	0.90	0.87	0.91	0.91	0.61	0.59	0.68	0.54
CLASS	CLPC	CORR	1.00	0.62	0.72	0.67	0.69	0.65	0.65	-0.02	-0.02	0.02	-0.08
CLASS	CLPC	ICC	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18
CLASS	CLPC	AdjICC	0.57	0.56	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
CLASS	CLNC	CK	1.00	0.26	0.29	0.22	0.38	0.27	0.27	NA	NA	NA	NA
CLASS	CLNC	QWK	1.00	0.41	0.49	0.39	0.47	0.37	0.37	NA	NA	NA	NA
CLASS	CLNC	ACC	1.00	0.77	0.82	0.73	0.84	0.80	0.80	NA	NA	NA	NA
CLASS	CLNC	ACC2	1.00	0.92	0.93	0.91	0.93	0.91	0.91	NA	NA	NA	NA
CLASS	CLNC	CORR	1.00	0.53	0.56	0.57	0.61	0.50	0.50	NA	NA	NA	NA
CLASS	CLNC	ICC	0.15	0.11	0.14	0.13	0.13	0.16	0.16	NA	NA	NA	NA
CLASS	CLNC	AdjICC	0.51	0.42	0.49	0.47	0.47	0.53	0.53	NA	NA	NA	NA
CLASS	CLTS	CK	1.00	0.25	0.35	0.34	0.36	0.21	0.21	NA	NA	NA	NA
CLASS	CLTS	QWK	1.00	0.55	0.61	0.68	0.67	0.54	0.54	NA	NA	NA	NA
CLASS	CLTS	ACC	1.00	0.46	0.53	0.53	0.56	0.43	0.43	NA	NA	NA	NA
CLASS	CLTS	ACC2	1.00	0.90	0.94	0.97	0.93	0.90	0.90	NA	NA	NA	NA
CLASS	CLTS	CORR	1.00	0.57	0.62	0.69	0.69	0.55	0.55	NA	NA	NA	NA
CLASS	CLTS	ICC	0.09	0.11	0.10	0.10	0.09	0.10	0.10	NA	NA	NA	NA
CLASS	CLTS	AdjICC	0.38	0.42	0.40	0.39	0.37	0.40	0.40	NA	NA	NA	NA
CLASS	CLRSP	CK	1.00	0.30	0.31	0.29	0.49	0.29	0.29	NA	NA	NA	NA
CLASS	CLRSP	QWK	1.00	0.63	0.66	0.72	0.74	0.65	0.65	NA	NA	NA	NA
CLASS	CLRSP	ACC	1.00	0.47	0.49	0.47	0.63	0.48	0.48	NA	NA	NA	NA
CLASS	CLRSP	ACC2	1.00	0.89	0.90	0.92	0.94	0.91	0.91	NA	NA	NA	NA
CLASS	CLRSP	CORR	1.00	0.64	0.67	0.74	0.74	0.66	0.66	NA	NA	NA	NA
CLASS	CLRSP	ICC	0.09	0.10	0.09	0.09	0.09	0.09	0.09	NA	NA	NA	NA
CLASS	CLRSP	AdjICC	0.37	0.41	0.39	0.38	0.38	0.38	0.38	NA	NA	NA	NA
CLASS	CLBM	CK	1.00	0.10	0.28	0.08	0.03	0.04	0.04	0.05	0.07	0.04	0.02
CLASS	CLBM	QWK	1.00	0.55	0.65	0.61	0.61	0.53	0.53	0.24	0.33	0.14	0.25
CLASS	CLBM	ACC	1.00	0.37	0.50	0.36	0.33	0.34	0.34	0.35	0.39	0.30	0.36
CLASS	CLBM	ACC2	1.00	0.86	0.89	0.86	0.91	0.86	0.86	0.85	0.90	0.76	0.89
CLASS	CLBM	CORR	1.00	0.63	0.71	0.70	0.73	0.64	0.64	0.29	0.39	0.19	0.30
CLASS	CLBM	ICC	0.25	0.23	0.25	0.24	0.25	0.26	0.26	0.24	0.24	0.24	0.25
CLASS	CLBM	AdjICC	0.66	0.64	0.66	0.65	0.66	0.68	0.68	0.66	0.66	0.66	0.66
CLASS	CLPRDT	CK	1.00	0.11	0.21	0.04	0.31	-0.02	-0.02	NA	NA	NA	NA
CLASS	CLPRDT	QWK	1.00	0.47	0.57	0.53	0.64	0.39	0.39	NA	NA	NA	NA
CLASS	CLPRDT	ACC	1.00	0.39	0.52	0.31	0.59	0.22	0.22	NA	NA	NA	NA
CLASS	CLPRDT	ACC2	1.00	0.83	0.84	0.83	0.90	0.83	0.83	NA	NA	NA	NA
CLASS	CLPRDT	CORR	1.00	0.62	0.66	0.73	0.76	0.60	0.60	NA	NA	NA	NA
CLASS	CLPRDT	ICC	0.07	0.07	0.07	0.07	0.07	0.08	0.08	NA	NA	NA	NA

Table 4: Full Agreement Metrics *(continued)*

Instrument	Item	Metric	Human	Encoder	un1	un2	un3	gte	e5	GPT	N	ND	NR
CLASS	CLPRDT	AdjICC	0.30	0.32	0.32	0.30	0.31	0.34	0.34	NA	NA	NA	NA
CLASS	CLILF	CK	1.00	0.25	0.35	0.24	0.33	0.13	0.13	NA	NA	NA	NA
CLASS	CLILF	QWK	1.00	0.57	0.68	0.63	0.73	0.52	0.52	NA	NA	NA	NA
CLASS	CLILF	ACC	1.00	0.46	0.53	0.46	0.65	0.35	0.34	NA	NA	NA	NA
CLASS	CLILF	ACC2	1.00	0.89	0.95	0.88	0.96	0.84	0.84	NA	NA	NA	NA
CLASS	CLILF	CORR	1.00	0.59	0.70	0.67	0.74	0.58	0.58	NA	NA	NA	NA
CLASS	CLILF	ICC	0.08	0.10	0.08	0.08	0.08	0.08	0.08	NA	NA	NA	NA
CLASS	CLILF	AdjICC	0.33	0.39	0.35	0.35	0.34	0.35	0.35	NA	NA	NA	NA
CLASS	CLCU	CK	1.00	0.27	0.35	0.32	0.31	0.29	0.29	NA	NA	NA	NA
CLASS	CLCU	QWK	1.00	0.64	0.70	0.63	0.73	0.63	0.63	NA	NA	NA	NA
CLASS	CLCU	ACC	1.00	0.47	0.50	0.50	0.49	0.51	0.51	NA	NA	NA	NA
CLASS	CLCU	ACC2	1.00	0.91	0.91	0.91	0.94	0.93	0.93	NA	NA	NA	NA
CLASS	CLCU	CORR	1.00	0.65	0.72	0.65	0.77	0.65	0.65	NA	NA	NA	NA
CLASS	CLCU	ICC	0.01	0.02	0.01	0.01	0.01	0.01	0.01	NA	NA	NA	NA
CLASS	CLCU	AdjICC	0.07	0.09	0.07	0.06	0.05	0.06	0.06	NA	NA	NA	NA
CLASS	CLAPS	CK	1.00	0.24	0.27	0.28	0.43	0.16	0.16	NA	NA	NA	NA
CLASS	CLAPS	QWK	1.00	0.58	0.63	0.70	0.69	0.56	0.56	NA	NA	NA	NA
CLASS	CLAPS	ACC	1.00	0.48	0.48	0.50	0.62	0.42	0.42	NA	NA	NA	NA
CLASS	CLAPS	ACC2	1.00	0.88	0.91	0.91	0.95	0.84	0.84	NA	NA	NA	NA
CLASS	CLAPS	CORR	1.00	0.59	0.66	0.73	0.70	0.60	0.60	NA	NA	NA	NA
CLASS	CLAPS	ICC	0.06	0.06	0.06	0.06	0.06	0.07	0.07	NA	NA	NA	NA
CLASS	CLAPS	AdjICC	0.27	0.28	0.29	0.29	0.28	0.30	0.30	NA	NA	NA	NA
CLASS	CLQF	CK	1.00	0.23	0.36	0.30	0.44	0.09	0.09	NA	NA	NA	NA
CLASS	CLQF	QWK	1.00	0.63	0.71	0.61	0.76	0.60	0.60	NA	NA	NA	NA
CLASS	CLQF	ACC	1.00	0.39	0.50	0.46	0.57	0.24	0.24	NA	NA	NA	NA
CLASS	CLQF	ACC2	1.00	0.84	0.89	0.87	0.93	0.78	0.78	NA	NA	NA	NA
CLASS	CLQF	CORR	1.00	0.64	0.71	0.61	0.77	0.63	0.63	NA	NA	NA	NA
CLASS	CLQF	ICC	0.06	0.07	0.06	0.06	0.06	0.06	0.06	NA	NA	NA	NA
CLASS	CLQF	AdjICC	0.27	0.31	0.28	0.27	0.27	0.29	0.29	NA	NA	NA	NA
CLASS	CLINSTD	CK	1.00	0.26	0.38	0.32	0.13	0.32	0.32	0.00	-0.01	0.02	-0.01
CLASS	CLINSTD	QWK	1.00	0.61	0.72	0.66	0.52	0.68	0.68	-0.01	-0.02	0.00	-0.01
CLASS	CLINSTD	ACC	1.00	0.45	0.53	0.47	0.38	0.49	0.49	0.15	0.13	0.21	0.10
CLASS	CLINSTD	ACC2	1.00	0.90	0.94	0.91	0.88	0.93	0.93	0.45	0.39	0.53	0.43
CLASS	CLINSTD	CORR	1.00	0.62	0.72	0.68	0.63	0.69	0.69	-0.02	-0.03	0.00	-0.04
CLASS	CLINSTD	ICC	0.08	0.09	0.09	0.09	0.08	0.09	0.09	0.09	0.09	0.08	0.09
CLASS	CLINSTD	AdjICC	0.35	0.37	0.37	0.37	0.35	0.37	0.37	0.36	0.36	0.36	0.36
CLASS	CLSTENG	CK	1.00	0.12	0.24	0.32	0.18	-0.01	-0.01	NA	NA	NA	NA
CLASS	CLSTENG	QWK	1.00	0.48	0.62	0.67	0.63	0.38	0.38	NA	NA	NA	NA
CLASS	CLSTENG	ACC	1.00	0.34	0.42	0.52	0.35	0.27	0.27	NA	NA	NA	NA
CLASS	CLSTENG	ACC2	1.00	0.81	0.92	0.91	0.88	0.69	0.69	NA	NA	NA	NA
CLASS	CLSTENG	CORR	1.00	0.57	0.65	0.69	0.73	0.53	0.53	NA	NA	NA	NA
CLASS	CLSTENG	ICC	0.06	0.07	0.06	0.06	0.07	0.07	0.07	NA	NA	NA	NA
CLASS	CLSTENG	AdjICC	0.28	0.30	0.29	0.28	0.29	0.30	0.30	NA	NA	NA	NA

F Disattenuated Correlations

Results for disattenuated correlations described in Section 5.3 and their confidence intervals are in Figure 5. Most items show correlated relationships after disattenuation and most with confidence intervals above 0.5, suggesting that the encoder models and the humans are likely identifying similar sources of underlying teacher variation for those items.

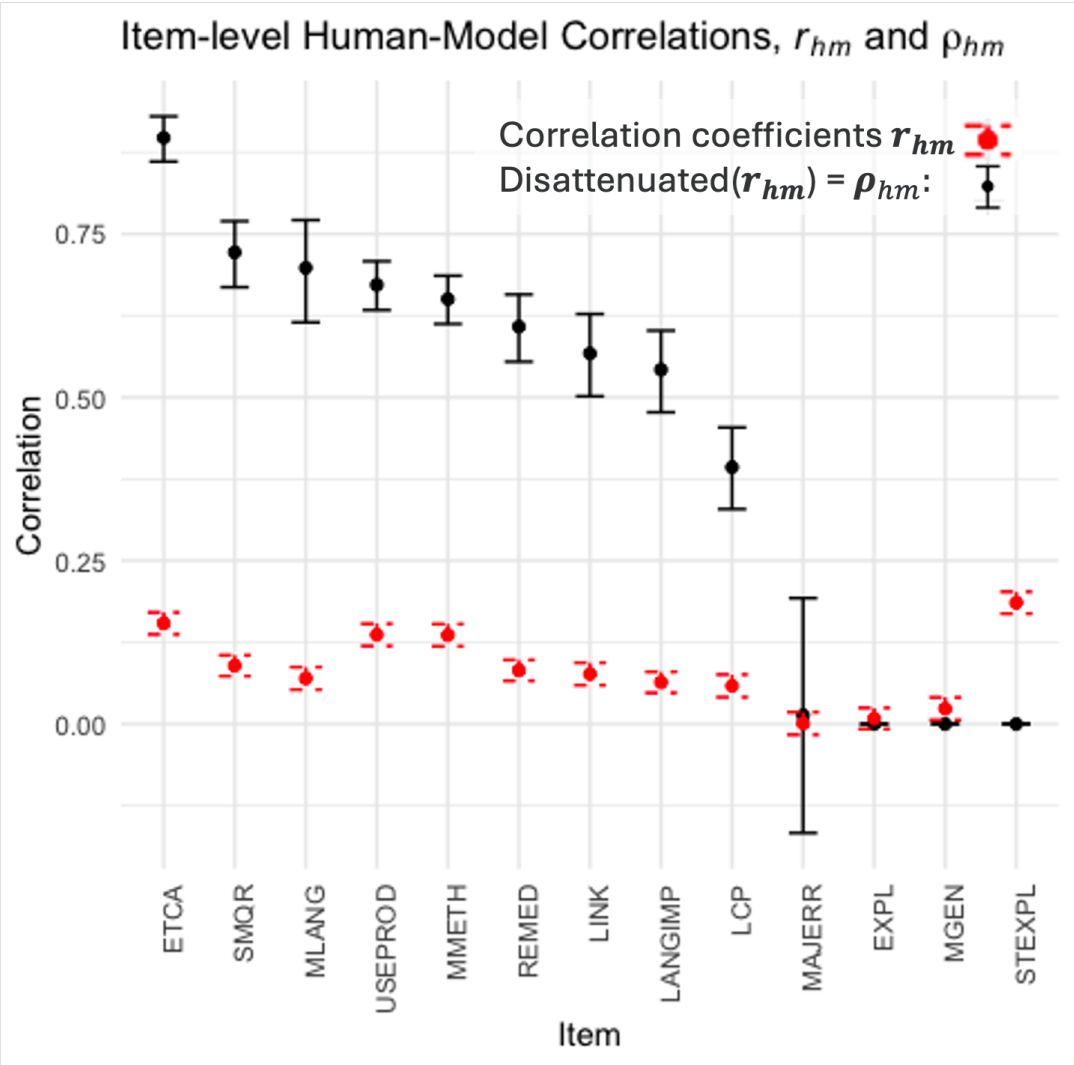

Figure 5: Correlations, Disattenuated Correlations and their respective 95% confidence intervals between human raters and model raters by MQI item. Item-level rater-label dependabilities for both human and model raters, Φ. The attenuated and disattenuated correlations between humans and models ρ_{hm} are shown. The attenuated correlation confidence intervals were calculated with the standard Fisher Transformation and $\alpha = 0.05$. Disattenuated correlation confidence intervals used the empirical method recommended in (Charles, 2005). GPT models are excluded as there were no correlations to disattenuate. Additional information about the disattenuation and reliabilities using the dependability metric, Φ, can be found in Appendix **??**

G Disentangling Bias and Measuring Fairness

Conducting a full fairness analysis across both CLASS and MQI items and raters is considerably more complicated when accounting for all four construct dimensions in (Blazar et al., 2017). If only MQI items are modeled, as was the case in the plots of Figure 1, the model can be simplified two dimensions. Full item-level MQI results for those models for disentangling biases from Section 5.4 are in Figure 6. The item-level results for corresponding racial bias difference models from Section 5.5 are in Figure 7. JAGS code for MCMC in R is available online.[17]

.5>

Listing 1: JAGS implementation for multi-instrument model

```
model
{
    for (i in 1:NN) {
        x[i] ~ dcat(prob.sdt[i, ])
        for (k in 1:K) {
            d[i, k] <- k - xi[
                subject[i], item[i]]
                - rhocov.r[rater[i],
                 item[i], race[i]]
            z[i, k] <- exp(-d[i, k]
                * d[i, k]/2 * exp(
                zeta.r[rater[i],
                 item[i], race[i]]))
            prob.sdt[i, k] <- ifelse
                ((K - maxscore.by.
                item[item[i]]),
                 ifelse(k < (maxscore
                     .by.item[item[i]]
                     + 1), z[i,
                   k]/sum(z[i, ]),
                       0.00000E+00), z
                       [i, k]/sum(z[i,
                   ]))
        }
    }
    for (i in 1:N) {
        for (j in 1:J) {
            xi[i, j] ~ dcat(prob.irt
                [i, j, ])
            for (m in 1:M) {
                kern[i, j, m] <-
                    alpha[j, m] * (
                    theta[i, m])
            }
            for (k in 1:K) {
                dotprod[i, j, k] <-
                    (k - 1) * sum(
                    kern[i, j,
                   ])
                eta[i, j, k] <-
                    dotprod[i, j, k]
                    - sum(gamma[j,
                  1:k])
                exp.eta[i, j, k] <-
                    exp(eta[i, j, k])
                prob.irt[i, j, k] <-
                     ifelse(K -
                    maxscore.by.item[
                    j],
                  ifelse(k <= (
                      maxscore.by.
                      item[j]), exp.
                      eta[i,
                    j, k]/sum(exp.
                        eta[i, j, 1:
                        maxscore.by.
                        item[j]]),
                    0), exp.eta[i, j
                        , k]/sum(exp.
                        eta[i,
                    j, 1:maxscore.by
                        .item[j]]))
        }
    }
}
for (nu in r1.raters) {
    for (s in r.1.in) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                ~ dnorm(eta.rt[s,
                 ra], prec.rhocov
                )
            zeta.r[nu, s, ra] ~
                dnorm(kappa.rt[s,
                 ra], prec.zeta)
            omega.r[nu, s, ra]
                <- sqrt(1/exp(
                zeta.r[nu, s,
               ra]))
        }
    }
    for (s in r.1.out) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                <- 0
            zeta.r[nu, s, ra] <-
                 0
            omega.r[nu, s, ra]
                <- 1
        }
    }
}
for (nu in r2.raters) {
    for (s in r.2.in) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                ~ dnorm(eta.rt[s,
                 ra], prec.rhocov
```

[17] to-be-added-after-anonymity-period

```
                )
            zeta.r[nu, s, ra] ~
                dnorm(kappa.rt[s,
                 ra], prec.zeta)
            omega.r[nu, s, ra]
                <- sqrt(1/exp(
                 zeta.r[nu, s,
               ra]))
        }
    }
    for (s in r.2.out) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                <- 0
            zeta.r[nu, s, ra] <-
                 0
            omega.r[nu, s, ra]
                <- 1
        }
    }
}
for (nu in r3.raters) {
    for (s in r.3.in) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                ~ dnorm(eta.rt[s,
                 ra], prec.rhocov
                )
            zeta.r[nu, s, ra] ~
                dnorm(kappa.rt[s,
                 ra], prec.zeta)
            omega.r[nu, s, ra]
                <- sqrt(1/exp(
                 zeta.r[nu, s,
               ra]))
        }
    }
}
for (nu in r4.raters) {
    for (s in r.4.in) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                ~ dnorm(eta.rt[s,
                 ra], prec.rhocov
                )
            zeta.r[nu, s, ra] ~
                dnorm(kappa.rt[s,
                 ra], prec.zeta)
            omega.r[nu, s, ra]
                <- sqrt(1/exp(
                 zeta.r[nu, s,
               ra]))
        }
    }
    for (s in r.4.out) {
        for (ra in 1:RA) {
            rhocov.r[nu, s, ra]
                <- 0
            zeta.r[nu, s, ra] <-
                 0
            omega.r[nu, s, ra]
                <- 1
        }
    }
}
for (m in 1:M) {
    pi.rt[m] <- 0
    delta.rt[m] <- 0
    sigma.rt[m] <- 1
}
for (s in 1:S) {
    for (ra in 1:RA) {
        eta.rt[s, ra] ~ dnorm(pi
            .rt[factors.by.item[s
            ]],
             prec.eta)
        kappa.rt[s, ra] ~ dnorm(
            delta.rt[factors.by.
            item[s]],
             prec.kappa)
        tau.rt[s, ra] <- sqrt(1/
            exp(kappa.rt[s, ra]))
    }
}
prec.pi ~ dgamma(a.precpi, b.
    precpi)
prec.delta ~ dgamma(a.precdelta,
     b.precdelta)
prec.eta ~ dgamma(a.preceta, b.
    preceta)
prec.kappa ~ dgamma(a.preckappa,
     b.preckappa)
prec.rhocov ~ dgamma(a.
    precrhocov, b.precrhocov)
prec.zeta ~ dgamma(a.preczeta, b
    .preczeta)
sd.rhocov <- sqrt(1/prec.rhocov)
sd.zeta <- sqrt(1/prec.zeta)
sd.pi <- sqrt(1/prec.pi)
sd.delta <- sqrt(1/prec.delta)
sd.eta <- sqrt(1/prec.eta)
sd.kappa <- sqrt(1/prec.kappa)
for (m in 1:M) {
    alpha[d2[1], m] <- ifelse(m
        == 2, 1, 0)
    alpha[d1[1], m] <- ifelse(m
        == 1, 1, 0)
    alpha[d3[1], m] <- ifelse(m
        == 3, 1, 0)
    alpha[d4[1], m] <- ifelse(m
        == 4, 1, 0)
}
for (j in d1[2:D1]) {
    alpha[j, 1] ~ dlnorm(0, prec
        .alpha)
    alpha[j, 2] <- 0
    alpha[j, 3] <- 0
    alpha[j, 4] <- 0
```

```
    }
    for (j in d2[2:D2]) {
        alpha[j, 2] ~ dlnorm(0, prec
            .alpha)
        alpha[j, 1] <- 0
        alpha[j, 3] <- 0
        alpha[j, 4] <- 0
    }
    for (j in d3[2:D3]) {
        alpha[j, 3] ~ dlnorm(0, prec
            .alpha)
        alpha[j, 2] <- 0
        alpha[j, 1] <- 0
        alpha[j, 4] <- 0
    }
    for (j in d4[2:D4]) {
        alpha[j, 4] ~ dlnorm(0, prec
            .alpha)
        alpha[j, 1] <- 0
        alpha[j, 2] <- 0
        alpha[j, 3] <- 0
    }
    for (j in 1:J) {
        gamma[j, 1] <- 0
        for (k in 2:maxscore.by.item
            [j]) {
            gamma[j, k] ~ dnorm(0,
                prec.gamma)
        }
        for (k in (maxscore.by.item[
            j] + 1):(K + 1)) {
            gamma[j, k] <- 0
        }
    }
    for (i in 1:TY) {
        for (m in 1:M) {
            ty[i, m] ~ dnorm(0, prec
                .ty)
        }
    }
    for (i in 1:N) {
        theta[i, 1:M] ~ dmnorm(ty[
            tyr.by.obs[i], ], Tau[,
            ])
    }
    Tau[1:M, 1:M] ~ dwish(W[, ], DF)
    Sigma <- inverse(Tau[, ])
    sd.th1 <- sqrt(Sigma[1, 1])
    sd.th2 <- sqrt(Sigma[2, 2])
    rho12 <- Sigma[1, 2]/sqrt(Sigma
        [1, 1] * Sigma[2, 2])
    prec.ty ~ dgamma(a.precty, b.
        precty)
    sd.ty <- 1/sqrt(prec.ty)
    prec.b <- pow(var.b, -1)
    prec.g <- pow(var.g, -1)
    prec.alpha <- pow(var.alpha, -1)
    prec.gamma <- pow(var.gamma, -1)
    prec.phi <- pow(var.phi, -1)
}
## initial values
inits <- function() list(
    alpha = item.dims * runif(J*M
        ,0.1,1.5),
    gamma = item.cats.by.score *
        rnorm(J*(K+1),0,0.5), #
    ty= matrix(rep(rnorm(TY, 0, 1),M
        ),nrow=TY,ncol=M),
    theta = matrix(rnorm(N*M,0,1),
        ncol=M),
    phi = rnorm(R, 0, 1),
    tau = runif(R, 0.1, 8),
    rhocov = array(rnorm(R*S*RA),dim
        = c(R,S,RA)) * rnorm(R*S*RA
        ,0,.5),
    zeta = array(rnorm(R*S*RA),dim
        = c(R,S,RA)),
    pi = rnorm(R,0,.5),
    delta = rnorm(R,0,.5),
    kappa = rnorm(R,0,.5),
    theta.prec = rgamma(1,100,100))
```

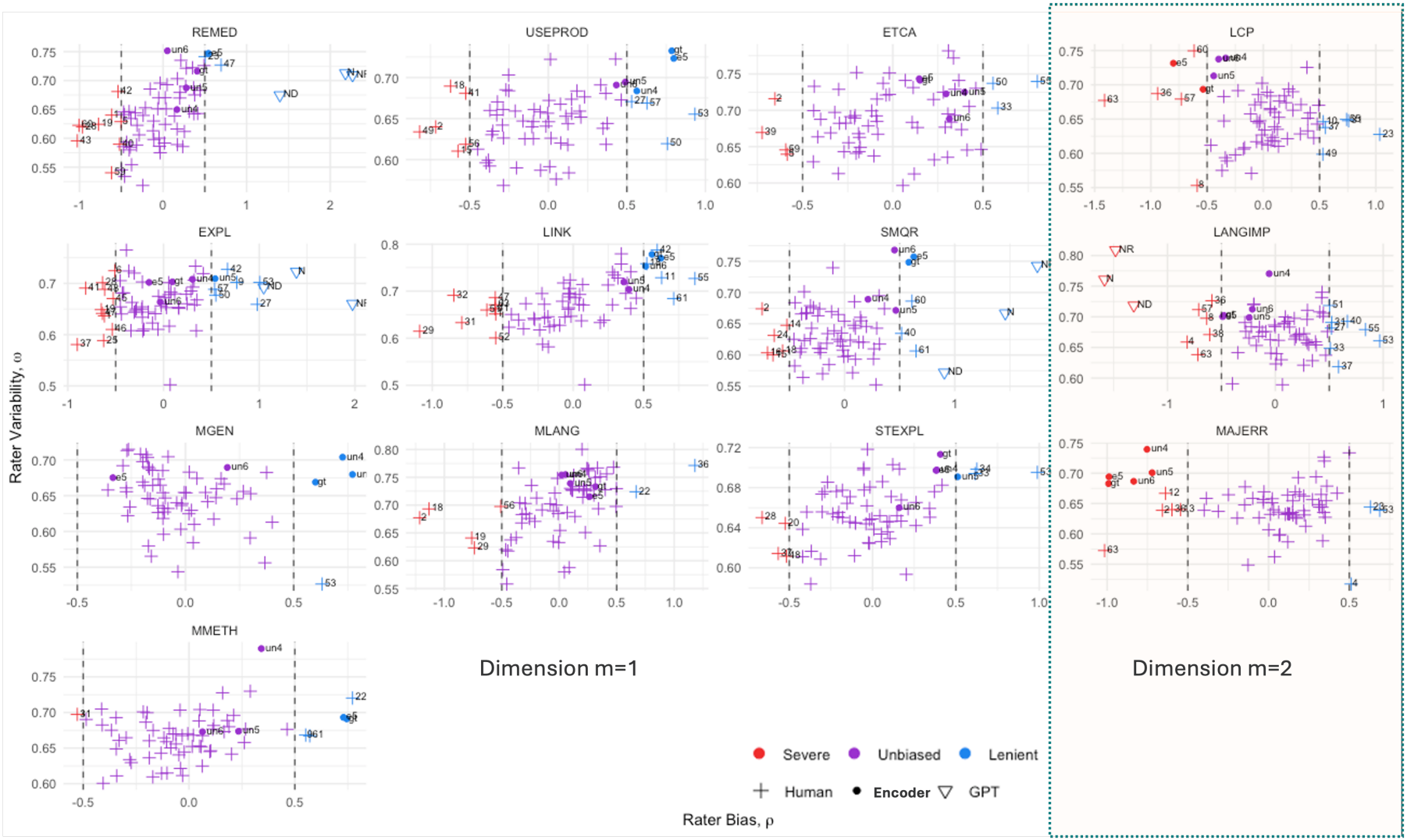

Figure 6: Rater biases, ρ_{jr} , for each item $j \in$ MQI centered at an item-level detection effectη_j, and variabilities, ω^2, by MQI Item and visually grouped by dimension, m and marked by severity/leniency. Each point is an individual rater: a "+" marker is a single human rater; "•" and "" are specific encoder and GPT models, respectively. X-axis is rater bias. Right is more lenient, left more severe. Color (via x-axis) are bias categories. Y-axis is rater variability (lower is more consistent. Horizontal lines 95% CI for bias via MCMC Bayes Estimation

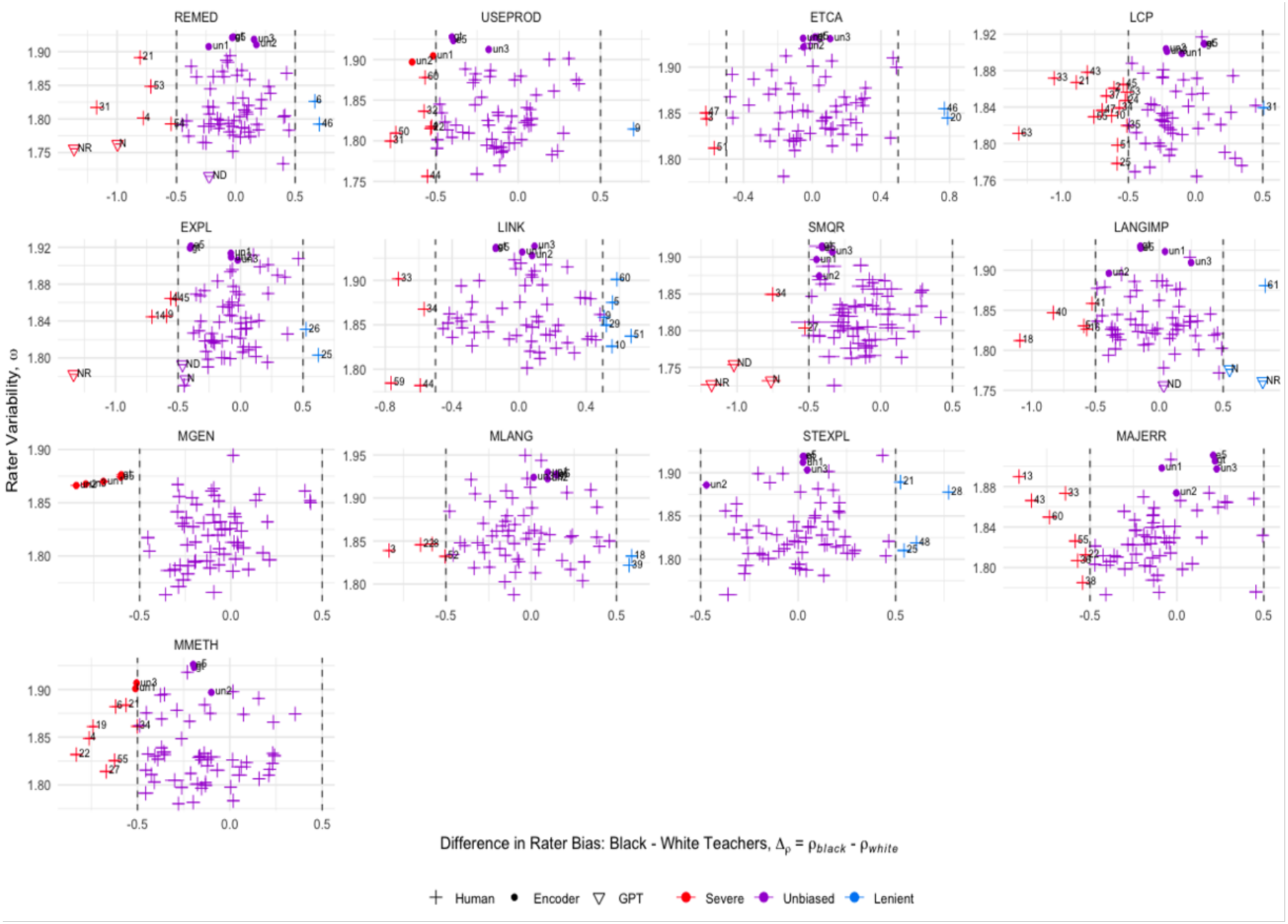

Figure 7: **Fairness across Racial Lines**. Section 5.5: Standardized difference in rater bias ϕ_r (x axis) and rater combined variability/consistency, ψ_r, (y axis) across Black teachers and White teachers. Leftward values are more severe towards Black teachers, rightward are more lenient. Any horizontal bar present with a marker represents 95% CI for bias. Differences in rater biases, $\Delta\rho_{jr}$, for each item $j \in$ MQI centered at an item-level detection effectη_j, and variabilities, ω^2 Each point is an individual rater: a "+" marker is a single human rater; "•" and "" are specific encoder and GPT models, respectively. X-axis is rater bias. Right is more lenient, left more severe. Color categories along x-axis are bias categories. Y-axis is rater variability (lower is more consistent. Horizontal lines 95% CI for bias via MCMC Bayes Estimation

H Generalizability and Decision Studies

H.1 Generalizability Study Human Results (for NCTE Main Study)

The results of the item-level G-study for human expert ratings, consisting of only the estimates for individual items using the NCTE Main Study data (Kane et al., 2015) to replicate Section 2.d from the Appendix. All calculations and representations are according to the design details listed therein.

In the Appendix of the NCTE study, the authors submitted a G-study on the MQI instrument, but not for data of the study: they provide a separate G-study of only eight (8) different middle school teachers teaching three (3) lessons each with only nine (9) raters, instead of the corresponding 317 NCTE Study teachers with an average 5.34 lessons each and 63 raters. For completeness, this paper conducts the G-study for the NCTE main study Appendix, Section 3, using the NCTE dataset. The full results of the human label G-study are in Table 5.

	Percent of Variance			height
ITEM (*I*)	**Teacher**	**Lesson**	**Rater**	**Tch×Rat**
	It	***Io:t***	***Ir***	**Itr**
LINK	12.3	39.6	6.5	4.3
EXPL	11.7	21.3	22.0	0.0
MMETH	12.7	51.0	3.9	4.3
MGEN	2.4	13.5	13.3	1.2
MLANG	5.7	29.8	17.3	0.0
REMED	11.2	25.3	12.4	0.0
USEPROD	18.1	18.9	13.9	9.4
MAJERR	6.9	25.7	8.3	14.3
LANGIMP	8.0	29.0	14.9	3.6
LCP	9.0	26.1	16.9	13.4
STEXPL	23.1	29.1	9.0	0.0
SMQR	12.8	24.4	11.2	1.2
ETCA	14.8	22.2	13.7	0.0

Table 5: By item, the percentage contribution, excluding the residual (which accounts for the remainder of the variance), of each variance component in the given MQI Item's R x (O:T) Generalizability Study

H.2 Generalizability Theory Parameters and Code

A helpful heuristic for understanding the mathematics of G-theory might be they are very computationally similar to hierarchical mixed effect models, where estimates of interest are found in variation of the random effects. The two code blocks represent by item $(O : I) \times R$ and $(S : O : I) \times R$ parameterizations, respectively, using variable names from the original dataset. The former replicates the methods used in (Hill et al., 2012b) and the Appendix Section 2.d of (Kane et al., 2015) to create Table 5 in Appendix section H.1, and was used in this study to calculate the family generalizability metrics in Section 5.2, including those used in Section 5.3. The latter is used for the decision studies described in Section 5.6. Studies were conducted using `lme4` (Bates et al., 2015) in `R` (Team)

Full results for item-level d-studies as defined in Section 5.6 are in Figure 8. .5>

Listing 2: G-theory lme4 parameterization

```
for (item in ITEMS){
    m[[item]] <- lmer(data = df|>
                    select(ITEM)|>
                    filter(R_TYPE ==
                        rater.type),
        formula = SCORE ~ (1|NCTETID
            _SCHOOLYEAR_SP/OBSID) +
                    (1|RATERID) +
                    (1|RATERID:NCTETID_
                        SCHOOLYEAR_SP)
}
```

Listing 3: Decision study lme4 parameterizationtabsize

```
for (item in ITEMS){
    m[[item]] <- lmer(data = df|>
                    select(ITEM)|>
                    filter(R_TYPE ==
                        rater.type),
        formula = SCORE ~ (1|NCTETID
            _SCHOOLYEAR_SP/OBSID/
            CHAPNUM) +
                    (1|RATERID) +
                    (1|RATERID:NCTETID_
                        SCHOOLYEAR_SP)
}
```

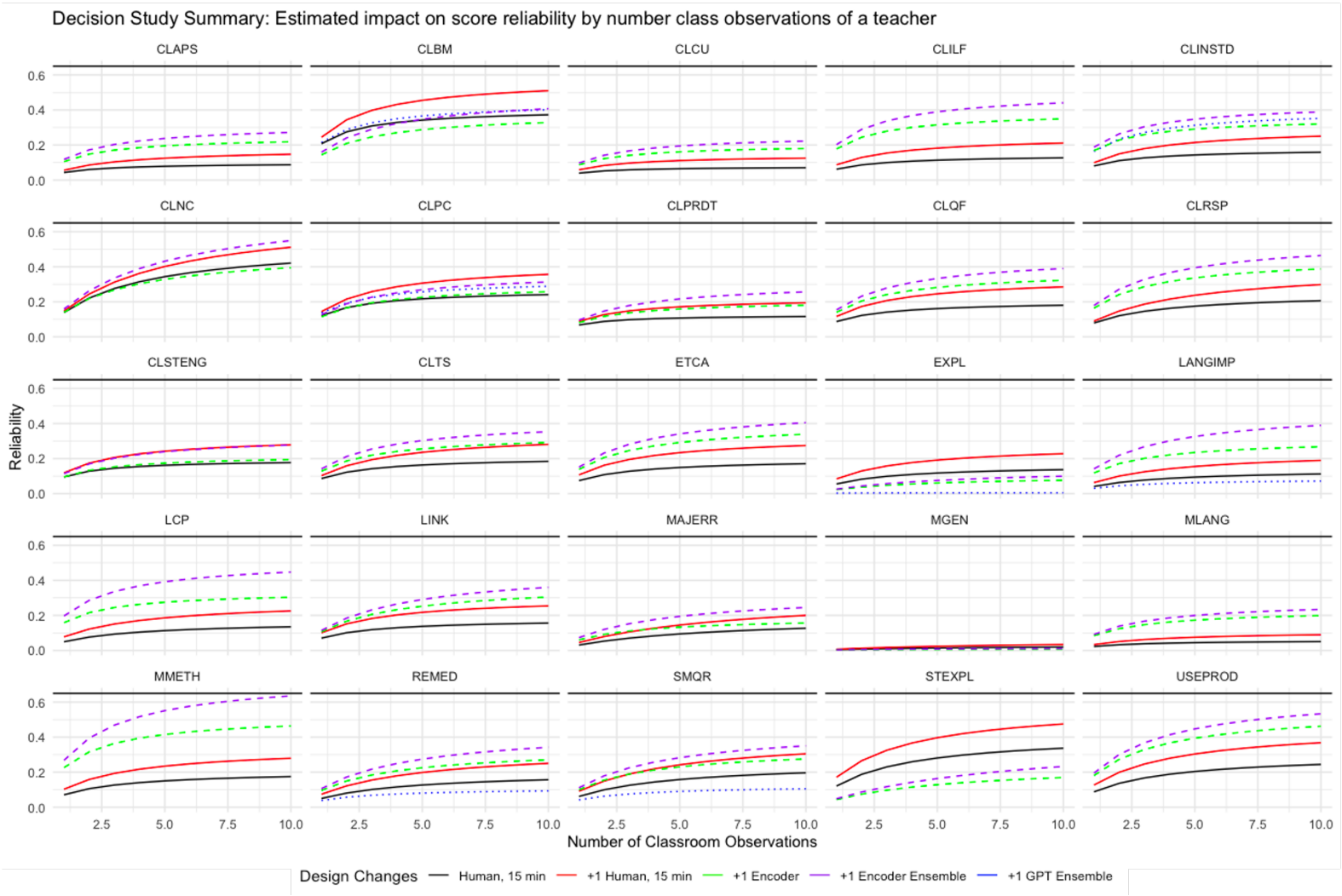

Figure 8: Expected changes to rating reliability are estimated improvements to quality (via reliability) of classroom ratings for various contexts. The single individual human baseline (black) estimates reliability improvements by visiting the same class the x axis represents the number of different 15 min. classroom observations of the same teacher. The red line is estimate of having a *different* human observer conduct observations as described. By contrast, for the model raters–single Encoder (green), Encoder ensemble (average of 3 encoders) (Red), and GPT ensemble (average of 3 GPT prompt engineered models)–the x-axis for models is the number of full classroom observations conducted where the human (black) observes at least 15 minutes (in-the-loop) of the same classroom (models observe the entire class period).

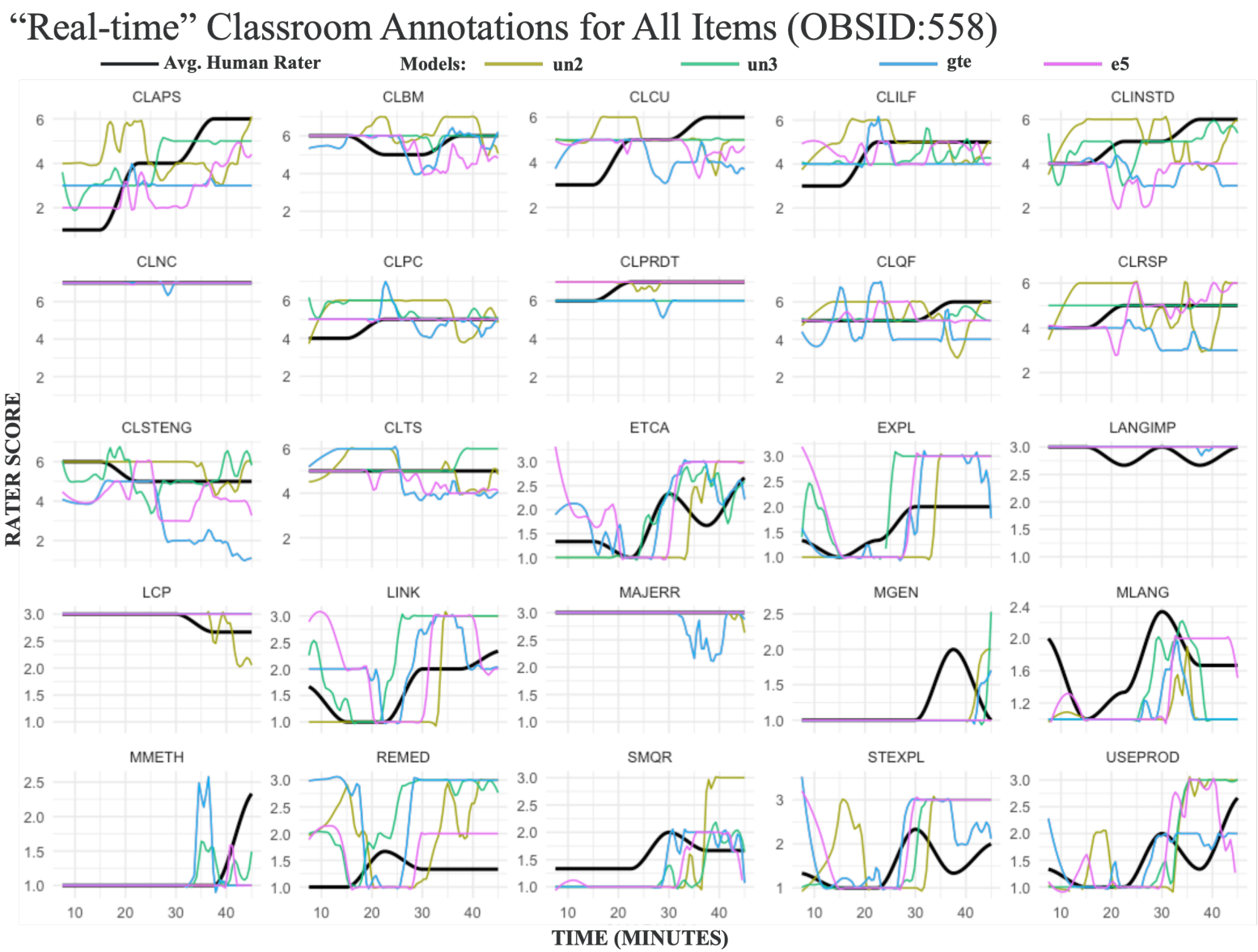

Figure 9: Real-time Evaluation: the X axis represents time in class (where 0 minutes is the start of class), each chart is one of the 25 items in the rubrics, the black lines are human evaluations (averaged, if multiple raters). The other lines are continuous model predictions for that item, using Loess smoothing where local fitting uses tricubic weighting of neighborhood points that span $\alpha = 0.1$.