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

Michael Hardy
Stanford University
hardym@stanford.edu

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 using standard metrics. However, evaluation techniques accounting for unreliable labels reveal important flaws, including spurious correlations and nonrandom racial biases across models and humans. We estimate that if models were used in a human-in-the-loop context, the variance contributed by GPT model labels would worsen ratings. These techniques also highlight tasks where encoders could offer 80% reduction in human costs while also reducing bias.

1 Introduction

Human-mediated labels always have an unknown amount of error. In machine learning practice, this error is often quantified using inter-rater reliability metrics and correlations. However, this annotation uncertainty is often ignored during standard supervised learning and model evaluation, leading to poorer models (Belz et al., 2023). Thus, imperfect labels are treated as "gold" or "ground truth" (Belz et al., 2020; Hosking et al., 2024). This may be due in part to the fact that accuracy measures are the most preferred methods of evaluating and benchmarking model performance (Birhane et al., 2022; Ribeiro et al., 2020; Kiela et al., 2021), but common practice could also arise from not using tools expressive enough to interpret labels in low reliability. To that end, this work demonstrates methods for working with low-/unknown-reliability annotations, often found in tasks requiring complex expert judgment.

The field of education has many complex tasks that often yield low reliabilities in labels (Jurenka et al., 2024; Kane and Staiger, 2012), which makes edtech NLP models and research particularly vulnerable to the effects of inexpert annotations (Belz et al., 2020; van der Lee et al., 2019; Zhou et al., 2023). The case study used to illustrate more expressive methods for working with unreliable labels will be from K12 education. Specifically, this study examines a use case where expert annotations are highly *unreliable* and yet *used in high-stakes decisions*: automated rating of **the quality of classroom teaching**. The methods used in this paper respond to the call of others to evaluate the psychometric properties of models that perform this task (Casabianca et al., 2013; Liu and Cohen, 2021), and do so by comparing metrics across six dimensions of interest: Concordance, Confidence, Validity, Bias, Fairness, and Helpfulness (complete results for these metrics compared to human baselines are in Table 2). In addition, novel contributions of this work to NLP include:

1. First demonstration of automated ratings of classroom instruction at or above the human-level reliability (Section 4.1, **Concordance**),
2. Measurements of the generalizability and dependability of labels found in NLP tasks (Section 4.2, **Confidence**),
3. Methods for detection of spurious correlations in model outputs (Section 4.3, **Validity**),
4. Methods for disentangling human rater-specific **biases** from data (Section 4.4) and

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2250–2278
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

measuring **fairness**, even in the presence of low label reliabilities (Section 4.5), and

5. Application of Design Studies (d-studies) from Generalizability Theory (g-theory) for estimating impacts of model use on human label quality (Section 4.6, **Helpfulness**).

This paper will explain the complexity of the case study task, followed by six evaluation methods to scrutinize the quality of the label. This work strengthens the argument that only using simple inter-rater reliability metrics to understand the quality of labels may mask the limitations of the labeling criteria (Hill et al., 2012b; Hosking et al., 2024; Belz et al., 2020). It also illustrates how more robust evaluation techniques can yield information in the presence of noisy labels and seemingly inconclusive results. The analyses presented in this study are motivated by issues of model interpretability, fairness, and usefulness (see Appendix A). Brief introductions to various techniques will be provided and illustrated via the study task, followed by discussion of the results.

1.1 Case Study: Rating Teaching Quality

The classification task of rating teaching 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 public education teachers in US K12 for both formative educator development feedback and as high-stakes teacher evaluations. Despite their ubiquity, these ratings are not reliable, even when conducted by experts (Ho and Kane, 2013; Kane et al., 2015; Kane and Staiger, 2012; Glaese et al., 2022; Whitehill and LoCasale-Crouch, 2024), similar to the poor reliability of other K12 education labels (Jurenka et al., 2024; Tack et al., 2023) that have limited the rigor of education research (Slavin, 2002; Klahr, 2013; Jurenka et al., 2024). Studies about ratings of instruction are also extremely expensive to conduct relative to other annotation tasks (Grissom et al., 2013; Liu and Cohen, 2021; Jurenka et al., 2024). There are only two major studies across hundreds of public school teachers that use authentic instructional metrics to support development: the MET study (Kane et al., 2013; Kane and Staiger, 2012) and the NCTE Main Study (Kane et al., 2015). The latter is the source of data for this study.

Ho and Kane estimated that increasing the number of human classroom observers can improve the reliability of assigned ratings. In their major work on the topic, they use methods similar to those in this paper to 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). Considering the expense, importance, complexity, and lack of reliability in ratings of classroom teaching and also the advances in natural language processing, automated ratings based on classroom discourse offer one potential solution.

Study Research Question: How can we know when the behaviors of models are good enough to be used lieu of humans as estimated by Ho and Kane when "gold" labels are not available?

Three recent studies have sought to use LLMs to provide classroom instruction ratings (via classroom transcripts) using authentic rating rubrics. Whitehill and LoCasale-Crouch (2024) provide instructional scores on a private dataset for Pre-Kindergarden classrooms using mixes of zero-shot LLMs and bag-of-words, commenting that their highest Pearson's correlation statistic of 72 experiments ($r = 0.48$) "approaches human inter-rater reliability". Wang and Demszky (2023) and (Xu et al., 2024) both use the same publicly available datasets as the present study, and the approach of the former will be discussed in more detail. Xu et al. use a by-item "best of" modeling approach and conducted many experiments with both zero-shot and fine-tuned encoder architectures (e.g., BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2020), etc.) and decoder architectures Llama 2 (Touvron et al., 2023) and ChatGPT. They report only the most performant models, greatly reducing the generalizability of the approaches. Although Xu et al. did not publicly release model ratings or the combinations of ensembles used, they reported Spearman correlation values for each of the best of several item-specific model constructions, which are also displayed in Figure 1.

2 Data

The data used in this study and in Wang and Demszky 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 6. Each item of the instrument is designed 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) item/task rating criteria, (2) raters/labelers, (3) stimuli/observations to be classified. As tasks increase in complexity, these facets contribute more error to estimates.

2.1 Rating Criteria: MQI Rubric

Imperfections in measurement instruments also add to measurement error. The 13 MQI items[2] within the dataset have at least two raters per classroom observation. Although both humans and encoders evaluated all items, this paper will focus on the 4 of the 13 MQI items[3] evaluated in Wang and Demszky (2023): teacher explanations (EXPL), remediation of student errors (REMED), student questioning and reasoning (SMQR), and imprecision in mathematical language (LANGIMP).[4] Additional information about the MQI instrument can be found in Appendices C and B.

2.2 Human Expert Raters

The 63 MQI raters[5] met a high standard: they were recruited from a separate pool of applicants based on their mathematics background and by contacting colleagues in mathematics departments (Hill et al., 2012a; Blazar et al., 2017), passed MQI certification exams, and attended biweekly calibration meetings to ensure standardization of scoring procedures.

2.3 Classroom Observations

Human raters watched videos and provided ratings on all MQI items at regular intervals. The 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 with human labels in the absence of timestamps. Data from the NCTE Main Study (Kane et al., 2015) [6] and for the associated transcripts[7] are available online.

3 Model Rater Families

GPT Models The GPT model family of Wang and Demszky (2023)[8] has 7,660 ratings for 223 different teachers. The family consists of three models that differ in prompt engineering methods, and a brief summary of these differences is given in Table 5. 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 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 4 and more training details can be found in Appendix D. In contrast to the GPT models, the only text pre-processing 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 family, all model outputs[9] used in evaluations and reporting in this study were conducted with a lesson-level-stratified held-out test set that

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

[2] instruments for classroom instruction are composed of multiple items, that represent distinct instructional dimensions to be evaluated

[3] Analyses for other items are in the appendices and online in Hardy (2024).

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

[5] 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.

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

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

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

[9] https://github.com/hardy-education/LLM-Psychometrics

was not used during model development. Encoder models were trained with a single GPU in Google Colab.

4 Evaluation Methods

4.1 Reliability and Concordance

RQ 1: How do automated raters perform relative to low-reliability labels?

Baseline Human Metrics Typical **reliability** metrics (see Section 4.1) provide a backdrop of descriptives based on **concordance** with other ratings. Full reproductions of all reliability metrics and calculation processes were performed exactly as described in the NCTE Main Study Appendix Section 2. (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. More details on the reproduced human results, model results, and additional metrics are in Appendix F.1.

Commonly Used Metrics Figure 1 shows the Spearman correlation ρ between the rater families. Common metrics also include: Pearson's correlation r, (e.g., Whitehill and LoCasale-Crouch, 2024), Kendall's correlation τ (e.g., Liu et al., 2023b) and Quadratic Weighted Kappa (QWK) typically used in ordinal classification tasks to penalize the distance quadratically (squared error) while accounting for categorical agreement by chance (e.g., Shermis, 2014; Hardy, 2021; Wang and Demszky, 2023). All these common label metrics for the four items in this study are in panel (b) in Figure 7 and the aggregated results by individual rating model for common metrics are in Table 6. Full reliability metrics at the segment level of lessons for all MQI and CLASS items can be found in Hardy (2024).

4.2 Confidence via Generalizable Reliability

RQ 2: How generalizable are findings based on unreliable labels?

Generalizability Studies (g-studies) (Brennan, 2001a, 2013, 2001b; Hill et al., 2012b) use random effect estimates across possible configurations of different sources of variance to quantify the stability of labels. This estimates the extent to which given labels 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 $\boldsymbol{\Phi}$, 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.

	$\mathbf{E}\hat{\rho}^2$		
ITEM	Human	Encoders	GPTs
EXPL	**0.15**	0.00	0.00
LANGIMP	0.09	**0.15**	0.08
REMED	**0.13**	0.10	0.05
SMQR	**0.14**	0.09	0.00
All Items	**0.114**	0.106	0.007

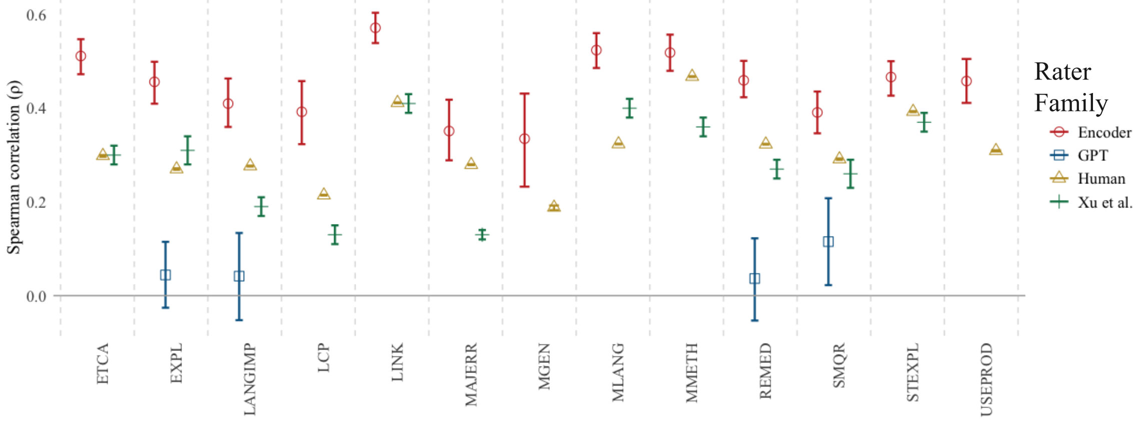
Table 1: Generalizability metrics by model families for each focus item. **Bold** represents the best rater family for $\mathbf{E}\rho^2$. In the overall "All Items" calculation, a $J \times R \times (O : I)$ model was used for comparability with other similar research. Generalizability ($\mathbf{E}\rho^2$) and dependability ($\boldsymbol{\Phi}$) results for all MQI items can be found in Table 8.

The multifaceted g-study design used to estimate how much variation 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 an Item-by-Rater-by-Segment-within-Observation-within-Individual Teacher design: $J \times R \times (S : O : I)$. The general estimates for all MQI items for a given rating family, $\mathbb{F}$, are shown in Table 8. For item-level reliabilities, we simplify the expression by keeping the item fixed, resulting in a $R \times (S : O : I)$ design. Using nested random effects notation, the estimation model is:

$$\begin{aligned} X^{(j)}_{s:o:ir} =& \mu + \nu_i + \nu_{o:i} + \nu_{s:o:i} \\ &+ \nu_{ir} + \nu_r + \nu_{s:o:ir} \end{aligned} \qquad (1)$$

where j indicates the item index, $\forall j \in \mathbf{J}$.[10] The code for the model specification is in Appendix I.4. Then, $\mathbf{E}\rho^2$ (Equation 2) and $\boldsymbol{\Phi}$ (Equation 11, Appendix I.4) are easily estimated from the random

[10] For the estimates in Fig. 7 (c), for dependability metrics of Section 4.3, 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 by-item $R \times (O : I)$ design, was conducted for the human expert rater family with results in Appendix I.1. The simplified model is $X^{(j)}_{o:ir} = \mu + \nu_i + \nu_{o:i} + \nu_{ir} + \nu_r + \nu_{o:ir}$ The full model structures of Eq. 1, 2 and 11 are used for Section 4.6.



Raters	ETCA	*EXPL*	*LANGIMP*	LCP	LINK	MAJERR	MGEN	MLANG	MMETH	*REMED*	*SMQR*	STEXPL	USEPROD
Humans	0.3	0.27	0.28	0.21	0.41	0.28	0.19	0.32	0.47	0.32	0.29	0.39	0.31
Encoders	**0.51**	**0.46**	**0.41**	**0.39**	**0.57**	**0.35**	**0.33**	**0.52**	**0.52**	**0.46**	**0.39**	**0.47**	**0.46**
GPTs		0.04	0.04							0.04	0.12		
Xu et al.	0.3	0.31	0.19	0.13	0.41	0.13		0.4	0.36	0.27	0.26	0.37	

Figure 1: Spearman correlation coefficients and confidence intervals by MQI Item for all rater families and studies. **Human** (Kane et al., 2015), **Encoder** (current study, Section 3), and **GPT** (Wang and Demszky, 2023) family correlations are between each rater and one randomly sampled human rater for each observation, following the processes used in the original human study, repeated 1,000 times for bootstrapped confidence intervals. **Xu et al. (2024)** coefficients are reported from Tables 5 and 9 of that paper, where each number represents the best of several ensemble models fit for each individual item. **Bold** in the table indicates highest performing label family. Italicized item abbreviations are those items evaluated by all studies.

effects for raters in rater family $\mathbb{F}$:

$$\mathbf{E}\rho^{2\,(j)}_{\mathbb{F}} = \frac{v_{ij}}{v_{ij}+v_{o:ij}+v_{s:o:ij}+v_{irj}+v_{s:o:irj}}, \qquad (2)$$

$\forall r \in \mathbb{F}$, where the item-rating-segment variation, $v_{s:o:irj}$, is confounded with the error variation. These results are found in Table 8.

4.3 Validity, Accuracy and Spuriousness

RQ 3: Do models and humans use similar observable features when annotating the same construct?

Disjoint Disattenuated Correlations Dependability and generalizability do not guarantee accuracy, but even at these very low levels, they can be used in indirect tests of convergent validity to see whether correlations between humans and models are low because of measurement error, such as poor rubric item construction, or because the two sets are really uncorrelated. Disattenuation does not change the low reliability across items nor the quality of the measurement, but it can offer evidence toward discerning model predictive validity by quantifying how changes in the underlying construct result in changes in the same direction for both human and model. If an individual teacher's latent instructional ability θ_i is about the same from lesson to lesson with the same students, we can correlate $\hat{\theta}_i$ for human ($\mathbb{h}$) and model ($\mathbb{m}$) family ratings for *different lessons* coming from the *same teacher* and correct for measurement error using each rater family's $\mathbb{F}$ label generalizability, $\mathbf{E}\hat{\rho}^{(j)}_{\mathbb{F}}$, for a given item j. The disattenuated correlation, $\rho^{(j)}_{\mathbb{hm}}$, between humans and a family of models for item, j, can be estimated:

$$\rho^{(j)}_{\mathbb{hm}} = \frac{\mathrm{Corr}[\tilde{\mathcal{X}}_{\mathbb{h}}(i,\mathfrak{L},j,r_{\mathbb{h}}),\tilde{\mathcal{X}}_{\mathbb{m}}(i,\neg\mathfrak{L},j,r_{\mathbb{m}})]}{\sqrt{\mathbf{E}\hat{\rho}^{2\,(j)}_{\mathbb{h}}\mathbf{E}\hat{\rho}^{2\,(j)}_{\mathbb{m}}}} \qquad (3)$$

where $\tilde{\mathcal{X}}_{\mathbb{F}}$ is score retrieval function for individual teacher i on item j by a random member r of rater family $\mathbb{F}$ in relation to some observed lesson $\mathfrak{L}$ with family label generalizability, $\mathbf{E}\hat{\rho}^{2\,(j)}_{\mathbb{F}}$ defined in Equation 2. The numerator of Eq. 3 is the correlation in scores whenever two different lessons from the same teacher were scored by raters from different families (human and model). The denominator then adjusts for based on the reliabilities of raters from each family to account for the known tendency of low reliability to diminish observed correlations.

4.4 Disentangling Sources of Bias

RQ 4: Can bias in individual rater behaviors be identified and disentangled from labels?

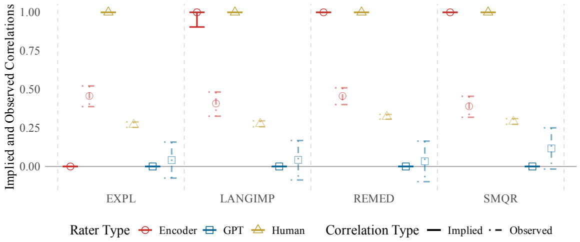


Figure 2: Observed (ρ, *fainter* color hues) and implied (ϱ_{hm}, **darker** color hues, Eq. 3) correlations between human raters and model raters by MQI item and their respective boostrapped 95% confidence intervals (1000 bootstraps of $N = 1500$ for both ρ and ϱ_{hm})

Hierarchical Rater Models 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, various polytomous IRT methods could help estimate "true/gold" labels (ξ_{ij}) during classroom observations, teacher instructional abilities (θ_i), and various effects of individual items. However, human raters introduce additional sources of measurement error for each rating and the data include multiples 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:

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

where an **IRT model** estimates the "gold" label score ξ_{soij} for a given item for some 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 that estimates the true scores based on the specific parameters of the rubric item and teacher is a multidimensional generalized partial credit model (MGPCM) (Muraki, 1992; Adams et al., 1997; Cui et al., 2024; Casabianca, 2021) with K_j categories. 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, $P\left[\xi_{oij} = \xi \mid \boldsymbol{\theta'}_i, \boldsymbol{\alpha}_j\ , \gamma_{j\xi}, o\right] =:$

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

where $oi = 1, ..., N$ lessons observed for teacher i, $j = 1, ..., J$ items, $r = 1, ..., R$ raters, and $k = 1, ..., K$ possible scores. Additional details on model specification and fit can be found in Appendix H.

As parameterized by Patz et al., the HRM base-level SDT model represents the measurement error induced by the rater r whose ability to "detect" the true score changes according to the individual rater's item-specific biases, ϕ_{jr} and variabilities, ψ_{jr} , on the x and y axes of Figure 3:

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

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 by Mariano and Junker.

4.5 Measuring Fairness across Racial Lines

RQ 5: With unreliable labels and complex tasks, can issues of racial fairness in ratings be disentangled from individual rater behaviors?

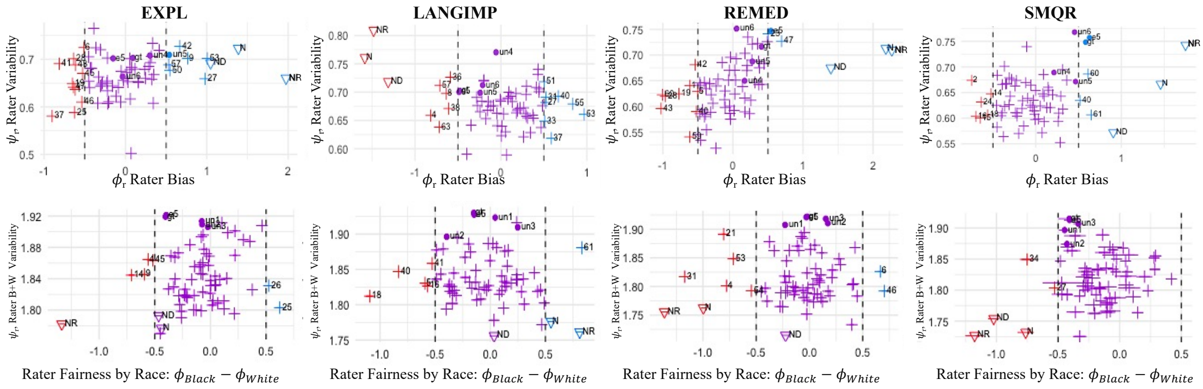


Figure 3: **(Top) Disentangled Rater Bias**. Section 4.4: standardized rater bias ϕ_{jr} (x axis) and rater variability/consistency, ψ_{jr} (y axis) from Equation 6, η_j-centered. Each point represents an individual human (represened by +) or model rater (Encoders and GPT models are •and ▽, respectively). More severe raters (**red**) are left, more lenient (**blue**) right. **(Bottom) Fairness across Racial Lines**. Section 4.5: Standardized difference in rater bias ϕ_r and rater combined variability/consistency ψ_r between ratings for Black teachers and White teachers. Leftward (**red**) values are more severe towards Black teachers (relative to White). Any horizontal bar present with a marker represents 95% CI from MCMC estimation. Rater estimates for additional items can be found in Hardy (2024).

Rater Covariate HRMs 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: $X \perp \varsigma \Rightarrow P_{\varsigma=a}(X_{jr}|\xi_j) = P_{\varsigma=b}(X_{jr}|\xi_j), \forall a, b$ which implies $\Delta\phi_{BW} = \phi_B - \phi_W = 0$. In the notation used for disentangling rater effects in Eq. 4, a score scoring rater r on item j is fair with respect to attribute ς given $\varsigma \perp \xi$:

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

The reparameterization of Equation 4 for rater covariates and additional details on model specification and fit can be found in Appendix H.

4.6 Estimating Real-world Helpfulness

RQ 6: Can we estimate the effects on rating quality and changes in real-world cost if a model were to be used with a human-in-the-loop?

Decision Studies (D-studies) estimate how reliabilities of ratings could improve by adjusting measured facets of variation, much like Ho and Kane did to motivate the case study. To estimate reliability in a human-in-the-loop scenario, multiple g-studies and d-studies would need to be constructed to combine variance contributions across a set of rater families, $\mathbb{F}$. For this work, only two different types of family are considered 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 of human confirmation. For a human-in-the-loop decision study, $\mathbb{F}$ would consist of families $\mathfrak{f}$ that have humans only and models only, and a combined human-model family. For a $(S : O : i) \times R$ study estimated dependability of ratings provided to teachers i on item j, $\widetilde{\Phi}_j$ is, in the joined "universe" $\mathbb{F}'$ where estimations are represented by $\mathbf{K}$, the collection of unique parameterizations and estimates, $\varkappa$, for the facets of variance in each D-study:

$$\widetilde{\Phi}_{j,\mathbb{F}'_{\varkappa} \sim \mathbf{K}} = \frac{\sum_{\mathfrak{f}}^{\mathbb{F}} \sigma^2(i_\varkappa)_{j\mathfrak{f}}}{\sum_{\mathfrak{f}}^{\mathbb{F}} \sigma^2(i_\varkappa)_{j\mathfrak{f}} + \sigma^2(\Delta_\varkappa)_{j\mathfrak{f}}} \quad (8)$$

where the summations in Equation 8 combines the variation across the familial "universes", indexed by $\varkappa$, of different rater families in $\mathbb{F}$ and $\sigma^2(i_\varkappa)_j$ and $\sigma^2(\Delta_\varkappa)_j$ represents the "universe" variability for teacher i and the absolute error for dependability, respectively, at the teacher-year-level (i) across the combined parameterization set $\mathbf{K}$. These values are represented in the ratio for calculating dependability, Φ_j, as found in Equation 11 $\sigma^2(\Delta)_j \equiv v_{o:ij} + v_{s:o:ij} + v_{irj} + v_{rj} + v_{s:o:irj}$. The expanded parameterization of $\sigma^2(\Delta_\varkappa)$ for the described human-in-the-loop model can be found in Appendix I.3.

5 Discussion

Table 2 shows a summary of the six dimensions of interest in the presence of low label reliability.

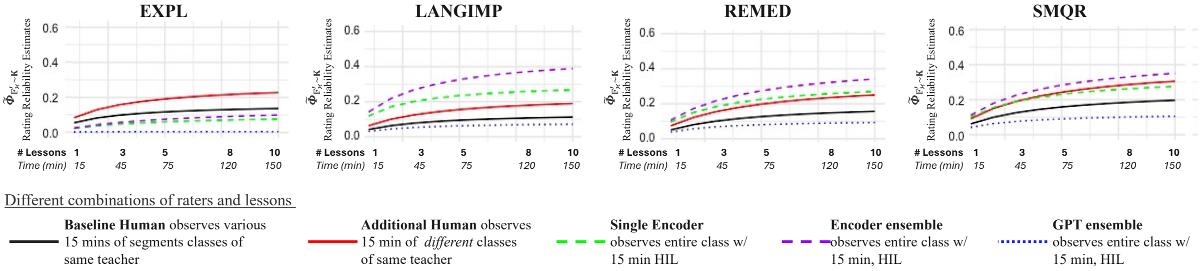


Figure 4: **Helpfulness: Estimated Improvements to Reliability**. Estimated improvements to the reliability of ratings under various mixes of humans, models, and numbers of lessons for the same teacher. The baseline is a single individual human observer (**black**, solid), showing modest improvements to estimated reliability with increased 15 minute visits to the teacher's classroom. The **red** (**solid**) represents the reliability resulting from a *different* human observer conduct additional observations (separate from the baseline observer). All model families only observe classes with the baseline observer: single Encoder (**green**, dashed), a 3-Encoder ensemble (**purple**, dashed), and GPT ensemble (3 separate prompts, **blue**, dotted). For models, the x-axis is the number of full classroom observations conducted where the human (black) observes a 15 minute portion of the same class.

	Category	Metric	GPTs				Encoders			
			EXPL	*LANGIMP*	*REMED*	*SMQR*	*EXPL*	*LANGIMP*	*REMED*	*SMQR*
RQ1	**Concordance**	*IRRs*	⊗	⊗	⊗	⊗	✓	?	✓	✓
		r, ρ, τ	⊗	⊗	⊗	⊗	✓	✓	✓	✓
RQ2	**Confidence**	$\mathbf{E}\rho^2$	⊗	⊗	⊗	⊗	⊗	✓	⊗	⊗
		Φ	⊗	⊗	⊗	⊗	⊗	✓	⊗	⊗
RQ3	**Validity**	$\varrho_{\text{hm}}^{(j)}$	⊗	⊗	⊗	⊗	⊗	✓	✓	?
RQ4	**Bias**	ϕ_r	⊗	⊗	⊗	⊗	✓	✓	✓	?
RQ5	**Fairness**	$\Delta\phi_{BW}$	?	⊗	⊗	⊗	✓	✓	✓	✓
RQ6	**Helpfulness**	$\widetilde{\Phi}_{\mathbb{F}'_{\text{HIL}}\sim\mathbf{K}}$	⊗	⊗	⊗	⊗	⊗	✓	✓	?

Table 2: Summary Performance Table for Families. **GPTs** are from Wang and Demszky and **Encoders** are from the present study. For each metric, symbols represent whether the model family generally performs as good as or better than humans ✓, worse than humans ⊗, or if performance relative to humans is unclear ?.

Figure 7 has the combined metrics from all sections. Using nearly any standard combination of metrics across all items, Encoder models perform better than even the *single highest performing expert human rater*. Additionally, the Encoder models' designs were constructed to allow for greater interpretability by evaluating continuous windows of classroom discourse, such as with real-time diagnosis and conducting sentence-level feature attribution[11] via integrated gradients (Sundararajan et al., 2017). However, their performance is not unidirectional: Sections 4.2 through 4.6 suggest that some SOTA-level correlations may have been spurious and provide insight into the poor performance of the GPT models.

Concordance Human raters meet an extremely high bar as annotators and yet show relatively weak correlations with each other on the MQI Instrument, corroborating similar findings from Kane and Staiger (2012). Despite this, the encoder family demonstrates that there is sufficient signal between the transcripts and human ratings to meaningfully perform the task of rating classroom instruction, ***outperforming human concordance in 100%*** of MQI tasks in Fig. 1. We suspect Encoder performance was higher than that of the GPT models, mostly because they learned to generalize across raters and lessons, and GPT models relied on in-context learning. As previously reported in Wang and Demszky (2023), the zero-shot GPT models show that they have "significant room for improvement" in these tasks.

Confidence Although the encoder family outperformed the individual human rater in terms of concordance, they do so with *less consistency at the teacher level* (Table 8). Trained human experts are

[11] The material online in Hardy (2024) contains examples of the models' performance on continuous predictions. Code for statistical models is also available online.

likely better at selectively filtering for evidence that represents more persistent features of the construct being measured, even if they do not align or miss evidence on a single rating. The lack of generalizability of the encoders on EXPL ($\mathbf{E}\hat{\rho}^2 = 0$) suggests that they identified speech patterns *associated with higher observation scores but not necessarily specific to the quality of teacher explanations*.

Validity When we correct for measurement error, human raters have an implied correlation of $\varrho_{hm} = 1.0$ on all items (Fig. 2). This means that if infinite human experts were to rate the same lesson, scores would converge. This is also mostly true for Encoder models.[12] The failure of disattenuation to identify viable human-model correlations for items that previously showed correlated relationships in Section 4.1 suggests that previous correlations may be spurious. Humans and encoders showed agreement after correcting for measurement error. Contrastingly, low disattenuated correlations for GPT models suggest that they were, in fact, not correlated.

Bias Disentangling sources of bias from human raters can support data curation and flagging human raters, even when limited information about the raters themselves is known. This can be important when the sources of variation between observations are complex. GPT models ($\bar{\phi} = 0.85$) likely performed poorly in part due to the prompt length (Liu et al., 2023a) and the out-of-distribution nature of primary 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 not completely random. Education technologists and EdTech enthusiasts should be wary of the ability of foundation models to do out-of-distribution tasks.

Fairness Disentangling racial bias reveals differential rater functioning across racial lines with a negative bias ($\bar{\Delta\phi}_{BW} = -0.62$) against Black teachers relative to White teachers in Figure 3. Potentially more precisely, the centrality of the GPT model ratings appeared to diminish when rating black teachers, adding evidence that 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 the relative lack of familiarity of the models with AAE (Rickford and King, 2016). These results should give pause to edtech developers relying on prompt-engineering of foundation LLMs, as subtleties in biases exist in very complex tasks.

Helpfulness As conducting actual human-annotated classroom observation ratings is immensely expensive, the decision study analyses of Section 4.6 offer methods for estimating the improvement gained by using a model or model family. Figure 4 shows, for example, that if a school administrator were to observe two classes with an encoder ensemble, the estimated improvements in label dependability would be the same as two school administrators *each* observing five *different* lessons–**a cost savings of 80%**. 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. Notably, using GPT models in these scenarios would worsen the reliability of human ratings.

5.1 Conclusion

As foundation models are increasingly deployed in complex contexts where evaluation of the quality of their performance may not be feasible, identifying performance gaps in cases of unreliable annotations will be increasingly important, especially when downstream tasks diverge more from model training (McCoy et al., 2023). This paper demonstrated some techniques to show that even when human reliabilities are low, meaningful insights can be obtained to understand and improve model construction and use. These techniques uncovered checkered performance in answering whether models are good enough to be used to diagnose instructional quality. We encourage researchers to publish results and data 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.

[12] Disattenuated correlations are not directly comparable to the correlation measures in Section 4.1. Reported disattenuated correlations of 1.0 do not imply perfect correlations. They can mean that the measurement error is not randomly distributed.

6 Limitations

These methods serve as a proof-of-concept for improving reliability in widespread and costly classroom evaluation tasks. Even though these models can outperform a human given many accepted metrics, much more analysis and technological development is needed to ensure their readiness for influencing high-stakes decisions. Despite being best in class, the encoder models should not be used in production in their current state, even with a human in the loop. 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.

The methods in this paper are not representative of the full scope of psychometrics or of the best possible implementation of available methods. Rather, they 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, a reasonable assumption with humans. But this assumption might not be true for LLMs or for all tasks. A few models were estimated alongside humans to demonstrate how differently they behave under this assumption, but this paper provides no evidence that model abilities and underlying constructs as perceived by LLMs would be normally distributed (e.g., latent constructs could follow multimodal distributions 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 could potentially help, (e.g., unipolar IRT models Huang and Bolt, 2023 for detection tasks). Additionally, more facets of variation could be incorporated for more precise estimates. For example, Equation 8 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.

Although some studies cited in Appendix A.2 seek to generalize findings across all classrooms, this cannot be done with the transcript data used in this work, as it consists only of fourth- and fifth-grade mathematics classrooms from the United States. 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. The encoder models could be improved through metalearning during training, so they could be more adaptive to new instructional rubrics and classrooms. Their current transferrability is limited by their training and architecture just as much as lack of data for this task limits more robust model generalization. Furthermore, the models have not been trained to work with automated transcription, as the transcripts process was done with humans. 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. Although 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[13] Similarly, biases and imperfections of the MQI instrument (see Appendix C) would likely propagate.

And finally, while not as severe as the GPT results, the encoder models did not avoid issues of racial bias. On the item with lowest correlations 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. Earlier studies had already suggested that humans (see Appendix F.1 and Hill et al., 2012b) could not identify the MGEN item. Thus, the reasons for this bias arising are likely to do with label sparcity: <1% of training set labels had the highest rating on MGEN, likely leading to overfit on a potentially biased sample. This underrepresentation in data is a microcosm of the poor alignment between an LLM's training and the downstream task that GPT exhibits on a more global scale.

[13] 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.

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.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. *arXiv preprint*. ArXiv:2106.15590.

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. Https://doi.org/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.

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 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. 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 1638–1653, Online. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Jason Grissom, Susanna Loeb, and Benjamin Master. 2013. Effective Instructional Time Use for School Leaders: Longitudinal Evidence from Observations of Principals. *Educational Researcher*, 42(8)(42(8)):433.

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].

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

Michael Hardy. 2024. "All that Glitters": Approaches to Evaluations with Unreliable Model and Human Annotations. *arXiv preprint*. ArXiv:2411.15634 [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.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Shirley Ren, Udhay Nallasamy, Andy Miller, Kwan Ho Ryan Chan, and Jaya Narain. 2024. Do LLMs "know" internally when they follow instructions?

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].

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. *arXiv preprint*. ArXiv:2309.16349.

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.

Jing Liu and Julie Cohen. 2021. Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, 43(4):587–614. Publisher: American Educational Research Association.

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].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [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, Jay Belsky, Nathan Vandergrift, Renate Houts, and Fred J. Morrison. 2008. Classroom Effects on Children's Achievement Trajectories in Elementary School. *American Educational Research Journal*, 45(2):365–397. Publisher: American Educational Research Association.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*. ArXiv:1910.01108.

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.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. *arXiv preprint*. ArXiv:2306.06941.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

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 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://doi.org/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.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The Promises and Pitfalls of Using Language Models to Measure Instruction Quality in Education. *arXiv preprint*. ArXiv:2404.02444 [cs].

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint*. ArXiv:1906.08237.

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].

Xiaofei Zhou, Christopher Kok, Rebecca M. Quintana, Anita Delahay, and Xu Wang. 2023. How Learning Experience Designers Make Design Decisions: The Role of Data, the Reliance on Subject Matter Expertise, and the Opportunities for Data-Driven Support. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, pages 132–143, New York, NY, USA. Association for Computing Machinery. Event-place: Copenhagen, Denmark.

A Related Work

A.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).

A.2 Use of LLMs in 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.

B 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 3. The distributions of scores across all items for all rater families are in Figure 6. 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.

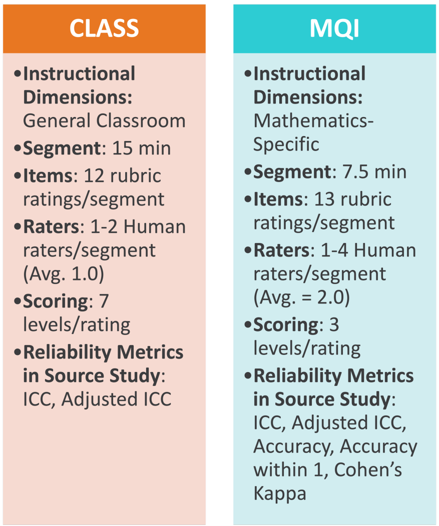

CLASS
•Instructional Dimensions: General Classroom
•Segment: 15 min
•Items: 12 rubric ratings/segment
•Raters: 1-2 Human raters/segment (Avg. 1.0)
•Scoring: 7 levels/rating
•Reliability Metrics in Source Study: ICC, Adjusted ICC
MQI
•Instructional Dimensions: Mathematics-Specific
•Segment: 7.5 min
•Items: 13 rubric ratings/segment
•Raters: 1-4 Human raters/segment (Avg. = 2.0)
•Scoring: 3 levels/rating
•Reliability Metrics in Source Study: ICC, Adjusted ICC, Accuracy, Accuracy within 1, Cohen's Kappa

Figure 5: Overview of technical details the two instructional frameworks used for evaluating instruction.

C MQI Instrument

C.1 MQI Instrument Properties

Previous 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 by reproducing the reliability metrics in Section 2.2, which correspond to the NCTE Study, Apdx Section 2). For our purposes, the MQI instrument has a few unique properties that warrant further analysis, since 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 in a single lesson.

The version of the MQI for which data are available in the NCTE dataset is ternary, in contrast to the current version of the MQI, which is quaternary. The lowest rating on the ternary MQI scale is a combination of the two lowest ratings on the quaternary, meaning that the present data cannot distinguish between whether the attribute described in each item is “Not present” or “Low”.[14] This ternary classification scheme creates non-normal distributions as seen in Figure 6, which will need to inform models and methods during quantitative analysis.

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

C.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.[15] 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 C.2 with further details.

Notably, 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 can be problematic, as it is asking for observers to use their

[14] 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.

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

Abbreviation	Item	Item Description
MQI		
ETCA	*Enacted Task Cognitive Activation*	Task cognitive demand, such as drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns.
EXPL	*Teacher Explanations*	Teacher explanations that give meaning to ideas, procedures, steps, or solution methods.
LANGIMP†	*Imprecision in Language or Notation*	Imprecision in language or notation, with regard to mathematical symbols and technical or general mathematical language.
LCP†	*Lack of Clarity in Presentation of Mathematical Content*	Lack of clarity in teachers' launching of tasks or presentation of the content.
LINK	*Linking and Connections*	Linking and connections of mathematical representations, ideas, and procedures.
MAJERR†	*Major Mathematical Errors*	Major mathematical errors, such as solving problems incorrectly, defining terms incorrectly, forgetting a key condition in a definition, equating two non-identical mathematical terms.
MGEN	*Developing Mathematical Generalizations*	Developing generalizations based on multiple examples.
MLANG	*Mathematical Language*	Mathematical language is dense and precise and is used fluently and consistently.
MMETH	*Multiple Procedures or Solution Methods*	Multiple procedures or solution methods for a single problem.
REMED	*Remediation of Student Errors and Difficulties*	Remediation of student errors and difficulties addressed in a substantive manner.
SMQR	*Student Mathematical Questioning and Reasoning*	Student mathematical questioning and reasoning, such as posing mathematically motivated questions, offering mathematical claims or counterclaims.
STEXPL	*Students Provide Explanations*	Student explanations that give meaning to ideas, procedures, steps, or solution methods.
USEPROD	*Responding to Student Mathematical Productions*	Responding to student mathematical productions in instruction, such as appropriately identifying mathematical insight in specific student questions, comments, or work; building instruction on student ideas or methods.
CLASS		
CLPC	*Classroom Positive Climate*	Positive climate reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and non-verbal interactions.
CLBM	*Behavior Management*	Behavior management encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior.
CLINSTD	*Instructional Dialogue*	Instructional dialogue captures the purposeful use of dialogue—structured, cumulative questioning and discussion which guide and prompt students—to facilitate students' understanding of content and language development. The extent to which these dialogues are distributed across all students in the class and across the class period is important to this rating.
CLNC†	*Classroom Negative Climate*	
CLTS	*Teacher Sensitivity*	
CLRSP	*Regard for Student Perspective*	
CLPRDT	*Productivity*	
CLILF	*Instr. Learning Formats*	
CLCU	*Content Understanding*	
CLAPS	*Applied Problem Solving*	
CLQF	*Quality of Feedback*	
CLSTENG	*Student Engagement*	

Table 3: CLASS and MQI item descriptions and the corresponding abbreviations. †denotes items that are reverse coded because they are negatively worded with respect to the construct of teacher ability. The bolded elements are those evaluated by the **GPT** rater family and reported by Wang and Demszky. Each member of the Human and Encoder families of raters evaluated all 25 items.

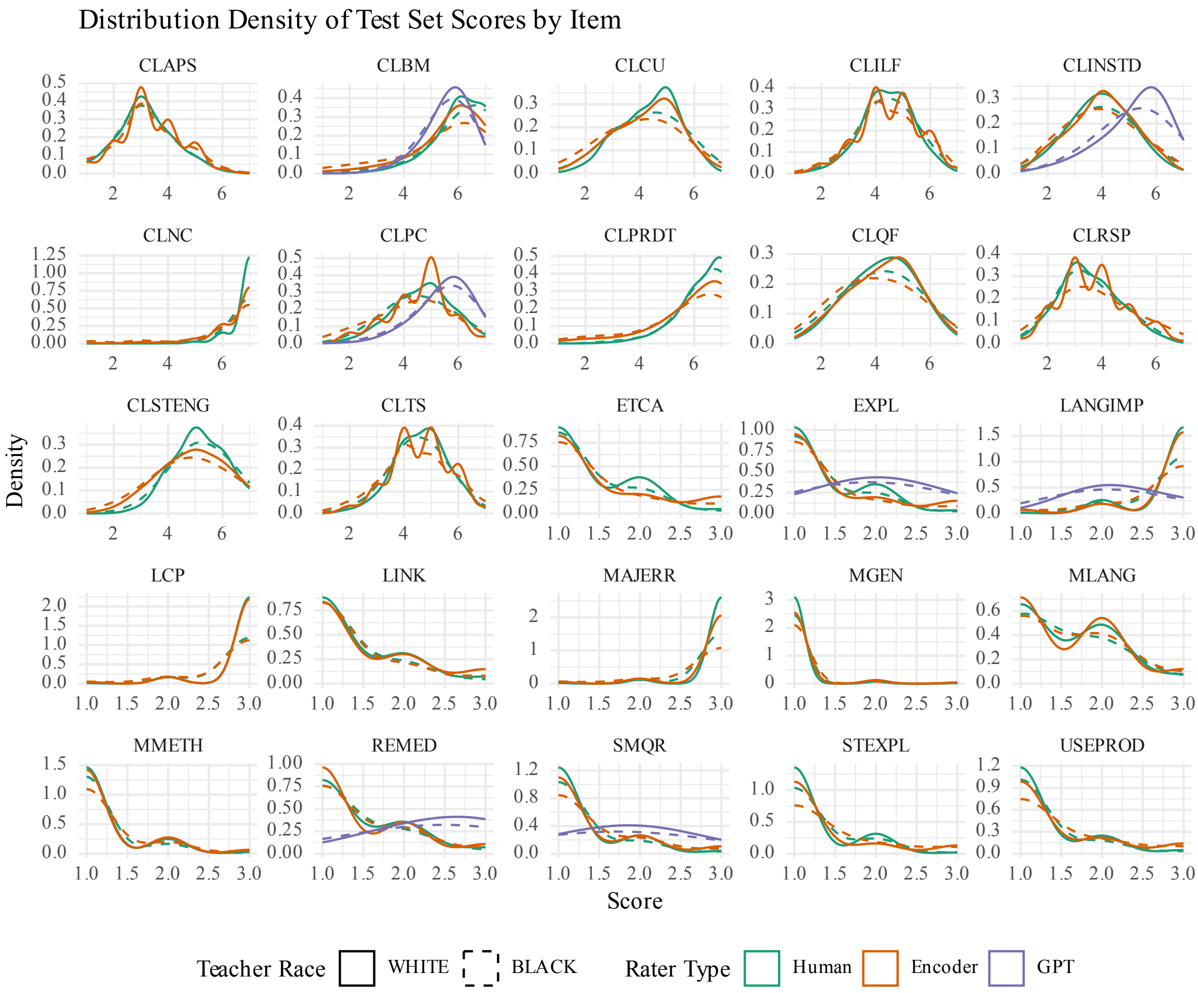
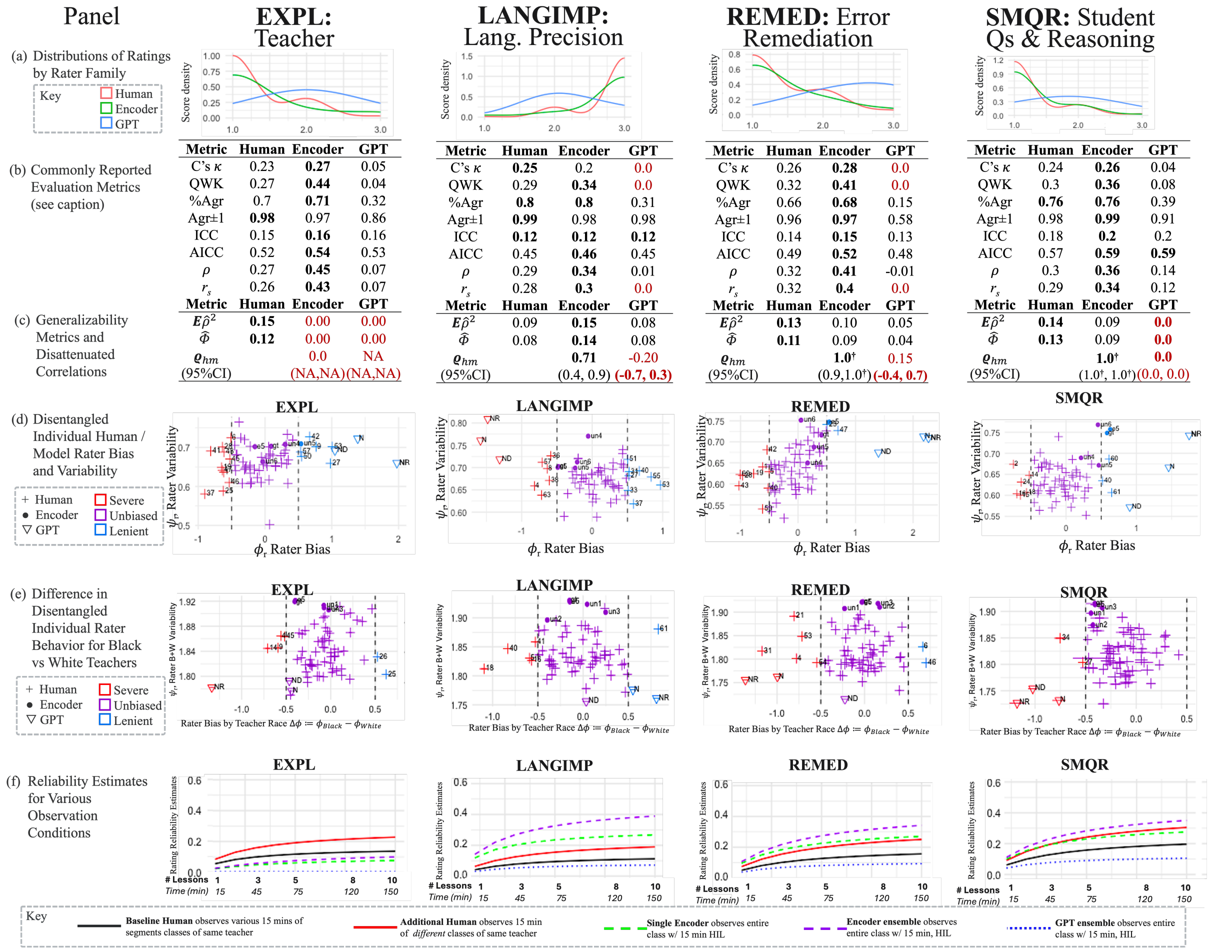


Figure 6: Distribution densities of rater scores for each of the 25 instrument items for all rater families.



EXPL: Teacher

Metric	Human	Encoder	GPT
C's κ	0.23	**0.27**	0.05
QWK	0.27	**0.44**	0.04
%Agr	0.7	**0.71**	0.32
Agr±1	**0.98**	0.97	0.86
ICC	0.15	**0.16**	0.16
AICC	0.52	**0.54**	0.53
ρ	0.27	**0.45**	0.07
r_s	0.26	**0.43**	0.07
Metric	**Human**	**Encoder**	**GPT**
$\boldsymbol{E}\hat{\rho}^2$	**0.15**	0.00	0.00
$\hat{\Phi}$	**0.12**	0.00	0.00
ϱ_{hm} (95%CI)		0.0 (NA,NA)	NA (NA,NA)

LANGIMP: Lang. Precision

Metric	Human	Encoder	GPT
C's κ	**0.25**	0.2	0.0
QWK	0.29	**0.34**	0.0
%Agr	**0.8**	**0.8**	0.31
Agr±1	**0.99**	0.98	0.98
ICC	**0.12**	**0.12**	**0.12**
AICC	0.45	**0.46**	0.45
ρ	0.29	**0.34**	0.01
r_s	0.28	**0.3**	0.0
Metric	**Human**	**Encoder**	**GPT**
$\boldsymbol{E}\hat{\rho}^2$	0.09	**0.15**	0.08
$\hat{\Phi}$	0.08	**0.14**	0.08
ϱ_{hm} (95%CI)		**0.71** (0.4, 0.9)	-0.20 **(-0.7, 0.3)**

REMED: Error Remediation

Metric	Human	Encoder	GPT
C's κ	0.26	**0.28**	0.0
QWK	0.32	**0.41**	0.0
%Agr	0.66	**0.68**	0.15
Agr±1	0.96	**0.97**	0.58
ICC	0.14	**0.15**	0.13
AICC	0.49	**0.52**	0.48
ρ	0.32	**0.41**	-0.01
r_s	0.32	**0.4**	0.0
Metric	**Human**	**Encoder**	**GPT**
$\boldsymbol{E}\hat{\rho}^2$	**0.13**	0.10	0.05
$\hat{\Phi}$	**0.11**	0.09	0.04
ϱ_{hm} (95%CI)		**1.0†** (0.9,1.0†)	0.15 **(-0.4, 0.7)**

SMQR: Student Qs & Reasoning

Metric	Human	Encoder	GPT
C's κ	0.24	**0.26**	0.04
QWK	0.3	**0.36**	0.08
%Agr	**0.76**	**0.76**	0.39
Agr±1	0.98	**0.99**	0.91
ICC	0.18	**0.2**	0.2
AICC	0.57	**0.59**	**0.59**
ρ	0.3	**0.36**	0.14
r_s	0.29	**0.34**	0.12
Metric	**Human**	**Encoder**	**GPT**
$\boldsymbol{E}\hat{\rho}^2$	**0.14**	0.09	**0.0**
$\hat{\Phi}$	**0.13**	0.09	**0.0**
ϱ_{hm} (95%CI)		**1.0†** (1.0†, 1.0†)	**0.0** (0.0, 0.0)

Figure 7: **Section 4 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 4.1. **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 4.2: generalizability coefficient $\mathbf{E}\rho^2$ and dependability measure $\boldsymbol{\Phi}$. Section 4.3: $\varrho_{\mathbf{hm}}$ is the disattenuated correlation. Red font indicates correlation was spurious or incalculable due to low reliabilities. **(d) Disentangled Rater Bias**. Section 4.4: 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 4.5: 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 4.6: 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).

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 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.[16]

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.

C.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 by combining G-theory and IRT estimations (Briggs and Wilson, 2007). However, the helpfulness of this study is 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 these 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.

D 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 4. Thus, the encoder family of models designed for this study shares the same architecture,[17] training, and held-out test sets, differing only as outlined in Table 4.

Another forthcoming paper explores this protocol in greater depth, showing that this training and augmentation of noisy data can similarly achieve SOTA and "superhuman" results on a variety of sentence embedding pre-trainings.

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.

D.0.1 Encoder Model Pre-processing

As mentioned in Section 3, pre-processing of the transcript data was intentionally mini-

[16] 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.

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

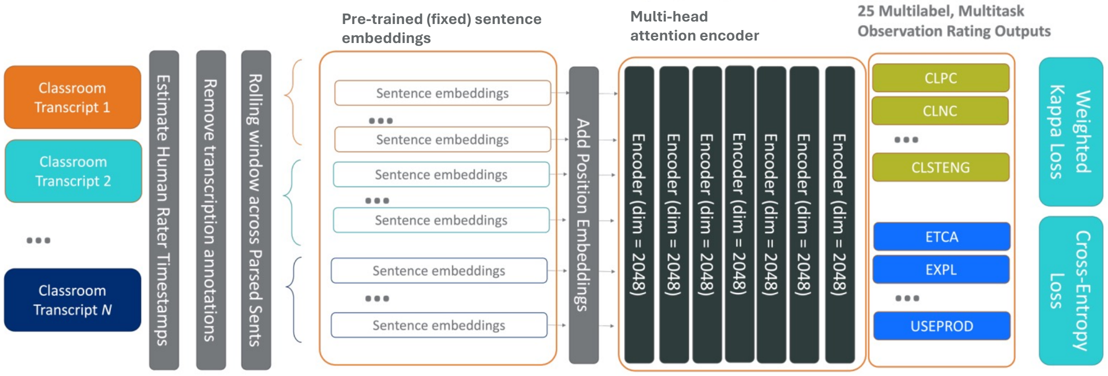

Pre-trained (fixed) sentence embeddings
Multi-head attention encoder
25 Multilabel, Multitask Observation Rating Outputs
Classroom Transcript 1
Classroom Transcript 2
•••
Classroom Transcript *N*
Estimate Human Rater Timestamps
Remove transcription annotations
Rolling window across Parsed Sents
Sentence embeddings
Add Position Embeddings
Encoder (dim = 2048)
CLPC
CLNC
CLSTENG
ETCA
EXPL
USEPROD
Weighted Kappa Loss
Cross-Entropy Loss

Figure 8: 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 4: Encoder Within-family differences: Summary of basic differences within the Encoder family of models. Detailed information on training and architecture can be found in the Appendix D.3.

mal, 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 further explored in (Hardy, 2024)

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.

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. Even though 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.

D.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

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

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

D.1.1 Embedding Model Selection

To save on compute, static embeddings were pre-computed. 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.

D.2 Model Architecture

D.3 Encoder Model Training and Description

Models were built and trained in pytorch,[18] 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.

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

- Attention heads: 32. Since a lot of semantic information were needed to be extracted from within each embedding and its neighbors, supporting an increase in multi-head self-attention mechanisms.
- Hidden dimension: 2048

D.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.

E GPT Model Family

E.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[19] which examples for how the three models differ. A brief summary of those differences can be found in Table 5.

E.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**).

F Reliability Metrics

As reproductions from the original NCTE study, ICC calculations were reproduced using the following multilevel model where lesson l scores for each rubric item are nested within teachers k:

$$ITEM_{lk} = \beta_0 + \mu_k + \varepsilon_{lk}, \tag{9}$$

and then calculate the ICC and Adjusted ICC

$$ICC = \frac{\text{var}\left(\mu_k\right)}{\text{var}\left(\mu_k\right) + \frac{\text{var}\left(\varepsilon_{lk}\right)}{n_l}}, \tag{10}$$

where $n_l = 1$ for ICC and where $n_l = 6$ for Adjusted ICC following the original study.

F.1 Full Results

Table 6 reports the relative performance on the metrics of Section 4.1. Full results of human baselines and comparisons against the various models can be found in the extended online material in Hardy (2024), which contains the all results calculations referenced in Section 4.1 and additional calculation details.

G Disattenuated Correlations

Results for disattenuated correlations described in Section 4.3 and their confidence intervals are in Figure 9. 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.

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

Metric	**Encoders**	*un1*	*un2*	*un3*	*gte*	*e5*	**GPTs**	*N*	*NR*	*ND*
%Agr	**0.54**	**0.69**	**0.77**	**0.69**	0.39	0.39	0.00	0.00	0.00	0.00
C's κ	**0.69**	**0.85**	**0.77**	**0.62**	**0.62**	**0.62**	0.00	0.00	0.00	0.00
QWK	**1.00**	**1.00**	**1.00**	**1.00**	**0.92**	**0.92**	0.00	0.00	0.00	0.00
r	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.00	0.00	0.00	0.00
ρ	**1.00**	**1.00**	**1.00**	**1.00**	**0.77**	**0.77**	0.00	0.00	0.00	0.00
τ	**1.00**	**1.00**	**1.00**	**1.00**	**0.77**	**0.77**	0.00	0.00	0.00	0.00

Table 6: **Concordance**: Performance above Human Reliability and Agreement Metrics. Proportion of MQI items where the *model* or **model family** listed had *better* results than human baselines. **Bold** indicates where performance was better on more than half of items rated. Inter-rater reliability metrics introduced in Section 4.1. **C's κ**: Cohen's κ; **QWK**: Quadratic Weighted Kappa; **%Agr**: percent exact agreement; r: Pearson's correlation; ρ: Spearman's rank correlation; τ: Kendall's concordance correlation;. Full data can be found in the supplementary material online.

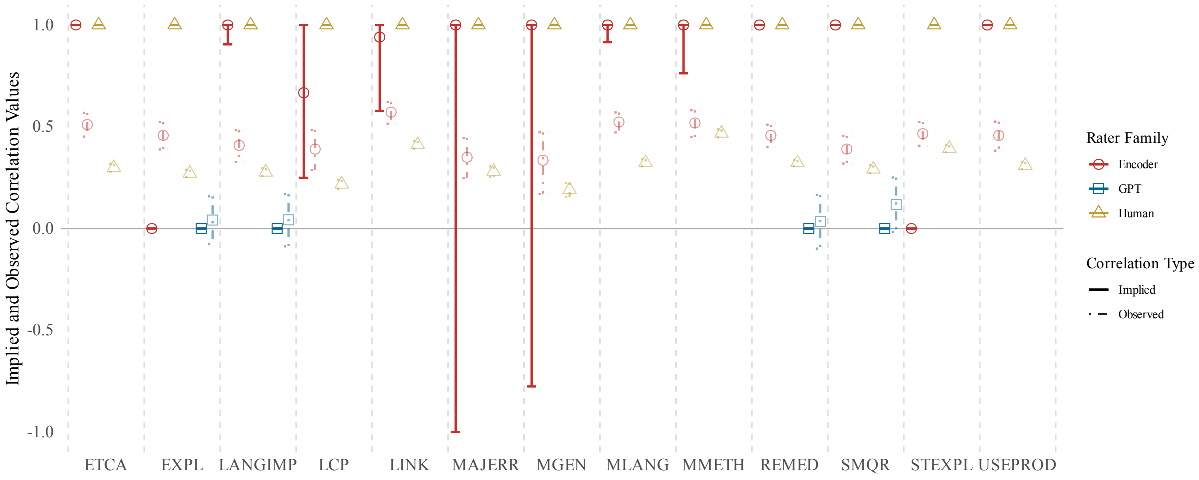


Figure 9: Correlations, Disattenuated Correlations and their respective 95% confidence intervals between human raters and model raters by MQI item. Observed (ρ, *fainter* color hues) and implied (ϱ_{hm}, **darker** color hues, Eq. 3) correlations between human raters and model raters by MQI item and their respective boostrapped 95% confidence intervals (1000 bootstraps of $N = 1500$ for both ρ and ϱ_{hm})

H Disentangling Bias and Measuring Fairness

H.1 Bias HRM Specification

Consequently, we update $\ln \psi^2_{jr} = \mathbf{Y}_{jr}(\ln \tau^2)$ where $\ln \tau^2 = (\ln \psi^2_1, ..., \ln \psi^2_R, \ln \tau^2_1, ..., \ln \tau^2_J)^T$. These results are in the top panel Figure 3. Bayesian estimates were calculated via Markov chain Monte Carlo (MCMC) simulation using Gibbs sampling across four chains with `JAGS` (Plummer, 2003) in `R` using very weakly informative priors and converging with $\hat{R} < 1.1$ for each parameter.

H.2 Fairness HRM Reparameterization

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 explicitly update the vector to illustrate the values $\eta = (\phi_{1_B}, ...\phi_{R_B}, \phi_{1_W}, ...\phi_{R_W}, \eta_{1_B}, ...\eta_{J_B}, \eta_{1_W}, ...\eta_{J_W})^T$ for the raters R, the items J, and $\ln \psi^2_{jr\varsigma} = \mathbf{Y}_{jr\varsigma}(\ln \tau^2)$ are updated similarly so that $\ln \tau^2 = (\ln \psi^2_1, \dots, \ln \psi^2_R, \ln \tau^2_1, ..., \ln \tau^2_J, \tau^2_B, \tau^2_W)^T$.

H.3 Fit and Code

For the models estimated in Section 4.5, less than 1% of the 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 3.

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 7, the model can be simplified two dimensions. This paper focused mostly on the MQI instrument, in particular the four items in the test set of Wang and Demszky. Item-level MQI results for the full model (including CLASS) for disentangling biases from Section 4.4 and for corresponding racial bias difference models from Section 4.5 with `JAGS` code listings, a model plate diagram and mixed effect model specifications are available in the extended supplementary material in Hardy (2024).

I Generalizability and Decision Studies

I.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 Kane et al. 2015 NCTE main study Appendix, Section 3, using the NCTE dataset. The full results of the human label G-study are in Table 7.

	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 7: 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

I.2 Rater Family Generalizability Metrics

The metric of dependability, Φ (Equation 11), is estimated from the random effects for raters in rater family $\mathbb{F}$ from the parameters in Equation 1:

$$\Phi^{(j)}_{\mathbb{F}} = \frac{\nu_{ij}}{\nu_{ij} + \nu_{o:ij} + \nu_{s:o:ij} + \nu_{irj} + \nu_{rj} + \nu_{s:o:irj}}, \tag{11}$$

Humans, on average, produce labels that are both more reliable and generalizable for capturing features that are more permanent at the teacher-level. The full results for human rater labels, decomposed into variance components, can be found in I.4[20] and estimates for $\mathbf{E}\rho^2$ and Φ can also be found in panel (c) of Figure 7. The Encoder models outperform humans on nearly every item in terms of inter-rater reliability metrics (Table 6) , but not in generalizable reliability metrics as seen in panel (c) tables of Figure 7. Importantly, the large difference between $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ for Humans and Encoders is due to properties of individual items, which accounted for over 75% of the variation in those families. GPT models, on the other hand, did not change ratings very much on different items, consistent with literature on these models not understanding such prompts (Liu et al., 2023a; Webson and Pavlick, 2022; Heo et al., 2024). Table 8 shows that Encoder model still performs better than humans on the majority of items, but it is no longer as clear. Interestingly, as mentioned in Section 3, the encoder models did not receive any annotations outside of the transcript, including speaker. This means that the model would struggle to identify teacher explanations (EXPL) from student explanations (STEXPL). This shift in interpreting encoder family performance from superhuman to zero reliability adds validity to the argument that these metrics provide valuable insight, showing that the relationships found in some of the variables could be explained by variance unrelated to the label construct. ***Implications***: Measures of generalizability and dependability derived from structured variance decomposition can meaningfully quantify label quality.

I.3 Decision Study Parameterization

Structurally, Equation 8 shares similarities with the two-stage ICC calculation (see Eq. 10). The absolute error for a rater family ($\mathbb{f}$) indexed by $\varkappa$ across any permutation of decision values in this study:

$$\sigma^2(\Delta_\varkappa) = \frac{\sigma^2(r_\varkappa)}{n'_{r_\varkappa}} + \frac{\sigma^2(o_\varkappa)}{n'_{o_\varkappa}} + \frac{\sigma 2(r_\varkappa i)}{n'_{r_\varkappa}} + \frac{\sigma 2(s_\varkappa : o_\varkappa)}{n'_{s_\varkappa} n'_{o_\varkappa}} + \frac{\sigma^2(s_\varkappa : o_\varkappa : ir_\varkappa)}{n'_{s_\varkappa} n'_{o_\varkappa} n'_{r_\varkappa}} \tag{12}$$

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 $\varkappa$, $n_{k_\varkappa}$ and those facets, if any, shared between families, $n_{k_{\mathbb{F}'}}$: $n'_{k_\varkappa} = n_{k_\varkappa} + n_{k_{\mathbb{F}'}}, \forall k \in \{s, o, r\}, n_{r_{\mathbb{F}'}} = 0$. These distinct sets of parameter values for each design study are represented in Equation 8. For human-in-the-loop *only* use cases, $\varkappa_{\text{HIL}}$, the value $n_{k_{\mathbb{F}'}}$ represents those sources of variation that are shared between rater families, and for a model family $\mathbb{f} = \mathbb{m}$, 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 n values where a human $\mathbb{h}$ observes a classroom for 15 minutes[21] with a model and where a single model $\mathbb{m}$ continues to observe for the remainder of the class (an additional 45 minutes), $\mathbf{K}_{n \in \varkappa_{\text{HIL}}} = \{n_{o_{\mathbb{m}}} = 0, n_{o_{\mathbb{h}}} = 0, n_{o_{\mathbb{F}'}} = 1, n_{s_{\mathbb{m}}} = 6, n_{s_{\mathbb{h}}} = 0, n_{s_{\mathbb{F}'}} = 2, n_{r_{\mathbb{m}}} = 1, n_{r_{\mathbb{h}}} = 1, n_{o_{\mathbb{F}'}} = 0\}$ and where the variance components are solved similarly to the coefficients of Eq. 1. Additional material can be found in Hardy (2024).

I.4 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 7 in Appendix section I.1, and was used in this study to calculate the family generalizability metrics in Section 4.2, including those used in Section 4.3. The latter is used for the decision studies described in Section 4.6. Studies were conducted using `lme4` (Bates et al., 2015) in `R` (Team). Full results for all 25 item-level d-studies as defined in Section 4.6 are can be found in Hardy (2024) as well as code listings used in the model estimations.

[20] Appendix 2.c of (Kane et al., 2015) provided a g-study, but, surprisingly, not using the data from the study.

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

		$\mathbf{E}\hat{\rho}^2$			$\hat{\Phi}$	
ITEM	Human	Encoders	GPTs	Human	Encoders	GPTs
ETCA	0.17	**0.20**		0.15	**0.19**	
EXPL	**0.15**	0.00	0.00	**0.12**	0.00	0.00
LANGIMP	0.09	**0.15**	0.08	0.08	**0.14**	0.08
LCP	0.11	**0.27**		0.09	**0.26**	
LINK	0.13	**0.19**		0.12	**0.19**	
MAJERR	**0.08**	0.00		**0.07**	0.00	
MGEN	0.03	**0.08**		0.02	**0.08**	
MLANG	0.07	**0.18**		0.06	**0.17**	
MMETH	0.13	**0.37**		0.13	**0.36**	
REMED	**0.13**	0.10	0.05	**0.11**	0.09	0.04
SMQR	**0.14**	0.09	0.00	**0.13**	0.09	0.00
STEXPL	**0.25**	0.00		**0.23**	0.00	
USEPROD	0.19	**0.25**		0.17	**0.25**	
All Items	**0.114**	0.106	0.007	0.010	**0.014**	0.004

Table 8: Generalizability and Dependability metrics by model families for each MQI Item. **Bold** represents the best rater family for each of $\mathbf{E}\rho^2$ and Φ, respectively. Underlined items are focus MQI items, because they were evaluated by (Wang and Demszky, 2023). For the overall "All Items" calculation, a $J \times R \times (O : I)$ model was used for comparability with other similar research.