PREDICTING GROWTH IN CLASSROOM DISCUSSION QUALITY WITH AUTOMATED SCORING

Benjamin Pierce, Rip Correnti, Nhat Tran, Diane Litman, Lindsay Clare Matsumura

University of Pittsburgh LRDC

PROBLEM AND RESEARCH QUESTIONS

High-quality classroom discussions have been linked to numerous positive educational outcomes (Applebee et al., 2003; Nystrand, 2006), and researchers have identified valid, reliable ways to score discussion quality. However, classroom discussions are often long and dense, and tracking learning across time in discussions often requires scoring a high volume of material. Recently, advances in artificial intelligence have allowed researchers to assess the quality of various learning materials, such as tests and essays, using automated scoring methods. Automated scoring allows researchers and educators to assess a high volume of student material in much less time than it would take human scorers. If similar methods could be adapted to automatically apply existing methods of scoring discussion quality to transcripts of classroom discussions, it would promote research into how educators learn to create high-quality discussions, as well as scale our ability to treat discussions as objects of learning assessment in themselves (Dale et al., 2022; Demszky et al., 2021).

In order to explore the suitability of AI-enabled automated scoring for assessing classroom discussion quality, we pose four research questions:

RQ1. What is the correlation between assignment of human-generated and automated scores across different dimensions of an existing discussion quality measure?

RQ2. What is the relation between overall discussion quality growth estimates when human scores are used when compared to automated scores?

RQ3. What is the relation between estimates for individual dimensions of discussion quality when human scores are used compared to automated scores?

FRAMEWORK AND STUDY

We draw data from a small-scale RCT on Content-Focused Coaching for improving discussion quality (Correnti et al., 2021). Teachers (n=62) were assigned to treatment or control conditions by stratified random sampling using percentage of English learners in the classroom to create sampling strata. After accounting for attrition and incomplete study participation, we formed an analytic sample of 31 teachers.

Treatment consisted of an online workshop, guided practice, and online 1:1 coaching using a literacy-coaching model developed at the Institute for Learning (Matsumura et al., 2012), as well as pre- and post-lesson coach-teacher meetings to discuss lesson plans and classroom video recordings, respectively. The study produced 3-6 classroom video recordings per teacher.

We use these recordings to construct a dataset of classroom discussions. We explore discussion quality using the *Instructional Quality Assessment*, a tool which has been extensively validated and used by educators and researchers (Matsumura et al., 2010, 2013). IQA scores are assigned to a classroom discussion through a rubric based on the number of times each *Analyzing Teaching Moves* (ATM) codes is used in that transcript (Correnti et al., 2015). Because prior work establishes the reliability of human ATM coding and the derived IQA scores for our discussion dataset (Correnti et al. 2021), we are in a position to answer our research questions by focusing on the IQA.

DATA AND METHODS

DATA AND CODING

We compared human and automated ratings of a corpus of classroom discussion transcripts. Transcripts were sourced from a randomized controlled trial of instructional coaching for discussion quality (Correnti et al. 2021).

- **112 classroom discussions**
- **31 classrooms (8 treated, 23 control; 18 4th grade, 13 5th grade)**
- **Diverse student sample (73% Hispanic, 61% low-income)**

A team of human raters coded classroom videos using the IQA and ATM codebook. Discussion codes were then transferred from videos to transcripts by matching timestamps. We then trained a natural-language processing (NLP) model (Tran et al. 2023) to predict IQA scores of individual discussions by assigning ATM codes to appropriate sequences of sentences in each transcript. We focus on four IQA dimensions that have very high reliability in human scoring as well as direct relevance to discussion quality. The single-measure intra-class correlation coefficient for human raters across all IQA dimensions ranged from .89 to .98. The human ICCs for the four IQA dimensions we attempted to predict automatically were:

1. ***Teacher links student contributions: .92***
2. ***Students link their contributions to other students': .96***
3. ***Students support their contributions with text-based evidence/explanations: .89***
4. ***Teacher presses students to say more: .90***

DISCUSSION QUALITY CONSTRUCT

We consider IQA dimensions 1-3 to form an aspirational discussion quality construct. Teacher linking, student linking, and student use of evidence/explanations are unreservedly valuable discussion components. Teachers pressing students to say more is considered transitional: it is more valuable than typical unambitious teacher questions, but a discussion featuring a high number of presses at the expense of 1-3 is considered to indicate a mid-level developmental stage.

We therefore measure classroom discussion quality by measuring growth in IQA dimensions 1-3. We use automated scoring to predict all four dimensions and examine correlations between human and automated scores for all four.

MODELS

We examine change over time in discussion quality using hierarchical linear models (Raudenbush and Bryk, 2002). Each IQA dimension is predicted using a hierarchical linear model, and we conducted each analysis for both human and automated scores. All models adjust for the same classroom-level demographic characteristics: percent of economically disadvantaged students, percent of female, Hispanic, African-American, Asian, and bilingual students, and prior-year reading and math scores. We also adjust for number of turns in each discussion, since this can affect the number of ATM codes assigned per discussion. Our unconditional model for each IQA outcome is as follows:

$$\mathbf{IQA_{ti} = \pi_{0i} + \pi_{1i}(Time_{ti}) + e_{ti}} \qquad \textbf{[equation 1.1]}$$

$$\mathbf{\pi_{0i} = \beta_{00} + \beta_{01}(Online\text{-}CFC)_i + r_{0i}}$$

$$\mathbf{\pi_{1i} = \beta_{10} + \beta_{11}(Online\text{-}CFC)_i + r_{1i}} \qquad \textbf{[equation 1.2]}$$

IQA_{ti}: IQA dimension score (or composite discussion quality score)

$\pi_{0i:}$ baseline IQA score for teacher i

$Time_{ti}$: change in time since baseline

π_{1i} : linear growth slope for teacher i

β_{00} : average baseline IQA score across teachers

$Online\text{-}CFC_i$: dichotomous indicator for teacher i for participation in treatment

β_{01}: difference in baseline IQA score for treated versus control teachers

β_{10} : average linear growth slope for teachers

β_{11}: difference in growth slope of IQA score for treated versus control teachers

e_{ti}: within-person residual

$r_{0i,}$ and r_{1i} are the between-teacher variance estimates for the intercept at baseline, and the linear growth slope, respectively

RESULTS

RQ1: HUMAN AND AUTOMATED SCORE AGREEMENT

- Correlations between human and automated scores for the same dimensions range from *r* = .85 to *r* = .91 (Table 1, lower left quadrant)
- Correlations between human scores for quality dimensions 1-3 are all r >.52 (upper left quadrant), while those between automated scores for those dimensions are approximately *r* = .30 (lower right quadrant)
- Exploratory factor analysis indicates that dimensions 1-3 form a unidimensional construct on both human and automated scores
- Only one set of scores shows no relationship: automated scores for dimensions 2(aspirational) and 4 (transitional)

Table 1. Correlations Among Human and Automated IQA Dimension Scores

		Human				Automated		
		Tch. Links Students	St. Links to Others' Ideas	St. Evd. and Expl.	Tch. Press	Tch. Links Students	St. Links to Others' Ideas	St. Evd. and Expl.
Human	St. Links to Others' Ideas	***.56***						
	St Evd. and Expl..	***.52***	***.84***					
	Tch. Press	*.45*	*.52*	*.67*				
Automated	Tch. Links Students	***.91***	.41	.41	.40			
	St. Links to Others' Ideas	.39	***.88***	.48	.34	**.30**		
	St. Evd. and Expl.	.43	.63	***.85***	.44	**.30**	**.31**	
	Tch. Press	*.42*	*.23*	*.57*	***.85***	*.32*	*-.00*	*.52*

Bold = discussion quality composite, bold/italic = human-automatic agreement

Table 2. Comparison between human-human and human-automated agreement

IQA Dimension	Human-human agreement	Human-automated agreement
Student links to others' ideas	.92	.91
Teacher links students' ideas	.96	.88
Student supports contributions with evd/expl	.89	.85
Teacher presses students to say more	.90	.85

RQ2: PREDICTING OVERALL DISCUSSION QUALITY GROWTH

We want to know whether we can make the same inference for the treatment effect on growth in overall discussion quality whether we are using human or automated scores. We estimate the effect of the intervention on the composite *overall discussion quality* (IQA dimensions 1-3) for the entire sample of classrooms, finding significant and positive growth coefficients using both human and automated scores.

Figure 1. Comparison of treatment effect on discussion quality growth estimates

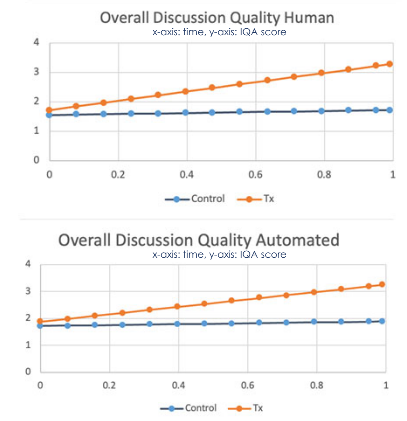

Table 3. T-statistics and effect sizes for treatment effect on discussion quality growth

	Human-scored transcripts	Machine-scored transcripts
T-statistic	3.60*	3.51*
Effect size	1.23	1.20

RQ3: PREDICTING GROWTH IN INDIVIDUAL DISCUSSION QUALITY DIMENSIONS

We want to know whether we can make the same inference for the treatment effect on growth across individual discussion quality dimensions whether we are using human or automated scores. We find that estimates of human- and automatically-scored transcripts are positive and statistically significant for two of the aspirational dimensions, and negative and statistically significant for the transitional dimension. The third aspirational dimension has mixed results: the growth estimate is positive and marginally significant for human-scored transcripts, but not significant for machine-scored transcripts. This dimension has the lowest incidence across transcripts of the four we consider.

Table 4. T-statistics for treatment effect on growth across individual IQA dimensions

	Human-scored transcripts	Machine-scored transcripts
Teacher links student contributions	1.07	1.08
Students link their contributions to others'	1.93	0.96
Students support their contributions with evd/expl	2.91*	2.66*
Teachers presses students to say more	-1.47	-1.61

* = p-value <0.5

CONCLUSIONS

KEY FINDINGS

- Automated scores reliably lead us to the same conclusions about the treatment effect in this RCT as human scores.
- Automated ATM codes are very closely correlated with human-assigned ATM codes (Table 1).
- Both human and automatic scoring of transcripts showed a positive, significant treatment effect (Figure 1, Table 2) on overall discussion quality growth.
- Automated scoring would lead to the same conclusion about growth as human scoring in three of the four individual IQA dimensions considered Table 3).
- Both human and automatic scoring found negative growth in the transitional IQA dimension. We theorize that this is consistent with the finding of growth in overall discussion quality because pressing students for further contributions is a transitional behavior, replaced in more advanced instruction by more specific forms of request and rejoinder.
- NLP methods can be used effectively to assess growth over time in the learning value of classroom discussions by automating existing validated methods.

LIMITATIONS AND FUTURE RESEARCH

- Human and automated scores differed on one of the IQA dimensions. Further work with our NLP model and others is necessary to understand why and improve performance.
- We are training our NLP model to apply more codes to more transcripts, including more sophisticated ways to understand student contributions.
- We are experimenting with different prompt engineering approaches to use large language models to assess discussion quality.

REFERENCES

Applebee, A. N., Langer, J. A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational research journal*, *40*(3), 685-730.

Correnti, R., Matsumura, L. C., Walsh, M., Zook-Howell, D., Bickel, D. D., & Yu, B. (2021). Effects of online content-focused coaching on discussion quality and reading achievement: Building theory for how coaching develops teachers' adaptive expertise. *Reading Research Quarterly*, *56*(3), 519-558.

Correnti, R., Stein, M.K., Smith, M.S., Scherrer, J., McKeown, M., Greeno, J., & Ashley, K. (2015). Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In L.B. Resnick, C.S.C. Asterhan, & S.N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 315–332). Washington, DC: American Educational Research Association.

Dale, Meghan E., et al. "Toward the automated analysis of teacher talk in secondary ELA classrooms." *Teaching and Teacher Education* 110 (2022): 103584.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.

Matsumura, L.C., Garnier, H.E., & Spybrook, J. (2012). The effect of Content-Focused Coaching on the quality of classroom text discussions. *Journal of Teacher Education*, *63*(3), 214–228. https://doi.org/10.1177/0022487111434985

Matsumura, L.C., Garnier, H.E., Slater, S.C., & Boston, M.D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*(4), 267–300. https://doi.org/10.1080/1062719080260 2541

Matsumura, L.C., Garnier, H.E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35–48. https://doi.org/10.1016/j.learninstruc.2012.11.001

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*, 392-412.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L. C. (2023, June). Utilizing natural language processing for automated assessment of classroom discussion. In *International Conference on Artificial Intelligence in Education* (pp. 490-496). Springer Nature Switzerland.

ACKNOWLEDGEMENTS & CONTACT

Benjamin Pierce: BEP51@pitt.edu

Work supported by a grant from the Learning Engineering Tools competition