IZA@LISER NETWORK

Discussion Paper Series

IZA DP No. 18942

September 2026

Eliciting Information about Teacher Applicants Using Open-Ended Questions

Joseph Aguilar-Bohorquez
University of Minnesota

Dan Goldhaber
University of Washington

Cyrus Grout
University of Washington

Elton Mykerezi
University of Minnesota

Prayash Pathak
University of Minnesota

Aaron Sojourner
W.E. Upjohn Institute for Employment Research and IZA@LISER

Eliciting Information about Teacher Applicants Using Open-Ended Questions*

Abstract

Employers commonly use open-ended questions in hiring, but usually responses aren't recorded, and knowledge about the usefulness of such responses is limited. This study investigates properties of written responses to open-ended questions collected over many years via an application tracking platform that school districts use to hire teachers. We develop a scalable process to discover topics within the corpus of responses to each question and then to classify each response with respect to the topics while preserving interpretability and transparency. It adapts a traditional qualitative topic modeling approach, Inductive Thematic Analysis, to best combine the strengths of researcher judgment and large language models (LLMs), introducing Hybrid LLM–Topic-Model Inductive Thematic Analysis (HLTM). Automated scoring achieves human-level reliability in document classification. Questions elicit distinct, non-redundant information regarding a candidate's pedagogical skills, classroom management, professional philosophy and growth mindset, relational and cultural competence as well as other professional attributes and institutional fit, but questions vary along important dimensions. Questions regarding classroom activities and student progress generate more dispersion in responses across applicants and fewer extreme demographic selection ratios. In contrast, philosophy-based questions produce more uniform, potentially scripted responses. This scalable framework empowers researchers and practitioners to assess language-based hiring instruments transparently and to monitor potential demographic disparities.

JEL classification

H75, I21, J45, D83

Keywords

education, hiring, teachers, topic modeling

Corresponding author

Aaron Sojourner
sojourner@upjohn.org

* This paper benefited from discussions with Brian Jacob, Lauren Dachille, Paul Sackett, Larisa Shambaugh, Beth Truesdale, and seminar participants at the Upjohn Institute and from funding from the U.S. Institute for Educational Sciences for project R305A220479-25. It was conducted under University of Western Michigan IRB Project Number 22-10-01.

1 Introduction

Employers commonly ask open-ended questions to help select among applicants for open positions. This type of question allow more expressive, personalized responses than other types of standardized assessments, potentially providing better information about applicants' skills, knowledge, and fit for a role (Posthuma et al., 2002; Macan, 2009). Digital and online hiring platforms have increased the collection of open-ended questions and artificial intelligence tools permit efficient evaluation of such questions, but empirical research quantifying the nature of the information derived from written responses to open-ended questions remains rare.[1]

We examine written responses to open-ended questions in the context of teacher applications. Although the methods developed and used below can be applied in many settings, teacher hiring is a particularly important context to research for two reasons. First, the public education sector is large and diffuse, employing more than three million teachers across thousands of school districts and other K–12 employers. Insights into a common hiring practice could have a broad impact. Second, teacher post-hire retention and effectiveness varies widely but is difficult to predict on the basis of easily-observed, pre-hire, administrative data Jackson et al. (2014). Bringing in additional information from psychological questionnaires and resumes has proven somewhat predictive Rockoff et al. (2011); Sajjadiani et al. (2019) but most variation remains unpredicted. Third, the quality of teaching and teacher hiring has clear ramifications for student outcomes and, thereby, impacts our nation's future productivity (Backes et al., 2024; Chetty et al., 2014; Jackson, 2018).

The framework developed here brings previously hard-to-observe information to light. Short answer questions are part of the decision maker's attempt to uncover some of this information pre-hire. To explore the quantification of open-ended questions, we draw on detailed data from a teacher applicant tracking system operated by Nimble Hiring PBC ("Nimble"). Nimble is used by numerous school districts and charter organizations to post vacant positions, collect applications, and screen and hire applicants. It provides employers an optional pre-populated bank of a dozen questions focusing largely on applicants' perspectives toward classroom practices, attitudes regarding student achievement, and approaches to interacting with students, families, and school administrators; employers can also craft their own questions. It is up to employers whether to require prospective employees to answer

[1]Questions are often asked and answered orally, and the information generated is often retained only in interviewer impressions or ratings rather than as an analyzable response corpus (Levashina et al., 2014). Even when interviews are technology-mediated, responses frequently remain oral (e.g., asynchronous video interviews), limiting availability of data collected in consequential, real hiring settings (Dunlop et al., 2022). More broadly, access to high-stakes organizational selection data is limited, reflecting legal and procedural constraints that make researcher access difficult (Potočnik et al., 2021; Langer et al., 2024).

any questions and how to use any answers as part of the applicant selection process.

This paper offers evidence on four research questions with respect to the dozen questions in Nimble's bank:

1. What topics do applicants tend to discuss in their responses and in what tone?
2. In scoring applicant responses into topics, how does automated scoring compare to human scoring?
3. Which questions produce responses that support more differentiation across applicants?
4. To what extent would selection on the basis of each question's response topics lead to adverse selection of applicants by race, ethnicity, and gender?

To measure applicant textual response, we innovate methodologically guided by the following three principles. First, maximize interpretability and transparency in the connection between the text and the topics, which is especially important in a practical hiring context. Second, maintain the critical role for researcher judgment and ensure researchers are well-informed about the text when making these judgments, as in standard Inductive Thematic Analysis (Braun and Clarke, 2006). Third, leverage automation in tasks where computers now have a comparative advantage in order to facilitate scalable, affordable analysis across a wide range of questions. We aim to characterize the topical content of responses in a way that is transparent, auditable, and intuitively meaningful to domain experts. This is essential in the context of hiring processes where applicants deserve fair treatment and an employer may have to explain a decision to a jury.

We develop Hybrid LLM–Topic-Model Inductive Thematic Analysis (HLTM), an approach that combines Inductive Thematic Analysis (ITA), large language models (LLMs), and topic modeling. We describe HLTM below, but first note that it is designed to address limitations of off-the-shelf approaches of textual analysis. Specifically, deductive approaches to textual analyses require the researcher to impose *ex ante* theory from the top down. This is defensible if the research team has expertise across all relevant content areas and a clear understanding of what to expect in the documents.[2] However, such top-down approaches are not suitable for analyzing a large, diverse set of documents that are not all well-theorized. [3]

[2]Examples of such top-down approaches include dictionary-based methods like LIWC (Linguistic Inquiry and Word Count), which quantifies text using pre-specified word-category dictionaries (Pennebaker et al., 2015), and supervised machine learning approaches that learn to apply researcher-defined categories from a human-labeled training set (e.g., Hopkins and King, 2010).

[3]Alternatively, fully automated inductive approaches such as Latent Dirichlet Allocation (LDA) are sensitive to nuisance parameters, such as the number of dimensions, and offer little transparency or interpretability. Similarly, inductive use of BERTopic and other LLM tools—ones that don't specify topics up front—has the same weaknesses.

ITA starts without top-down theory. Its initial stages generate evidence for a research team about prevalent meanings and empower the team to make well-informed decisions about defining a coherent set of topics that animates the underlying textual corpus. ITA is a widely used research tool. The original Braun and Clarke (2006) article that described it in a standardized way is one of the most cited works in the social sciences (almost 328,000 citations as of July 2026) and established thematic analysis as a foundational qualitative method. In personnel selection research, it is used in studies of recruiter decision-making, candidate evaluation, and bias, including analyses of social media influences (Melão and Reis, 2021) and the weighting of academic versus real-world experience (Donaldson, 2024). In education research, it is widely used to examine student experiences, curriculum design, and teacher perspectives, such as online learning during COVID-19 (Yeung and Yau, 2022), teacher views on student well-being (Byrne and Carthy, 2021), and assessments of problem-based learning.

Conventional ITA is costly to execute and scale, hence we explore how to combine LLM tools and researcher judgment within the ITA process. We focus research attention where human judgment has a comparative advantage and use LLM tools where their comparative advantages are. We automate early steps in the ITA process that involve summarizing chunks of text into phrases that resemble keywords (data coding in traditional ITA), then use embeddings-based topic modeling to infer dozens of higher-level themes prevalent in the text. As in traditional ITA, we decide the final set of topics, conceptual distinctions between them, and coding instructions for classifying text in the topics. It is this combination of ITA and LLM tools that we call Hybrid LLM–Topic-Model Inductive Thematic Analysis (HLTM). The HLTM process aims to remain inductive in nature by avoiding the imposition of a top-down structure on the data, while also ensuring transparency: each topic is traceably rooted in specific text from the documents, and each step from text to topic can be audited by humans or LLMs. The paper walks through the method with reference to a sample question (about classroom management). We apply the same method to every other question, reporting the resulting topics in the appendix.

Once topics are identified and described, we leverage LLMs to classify each response document in its question's topic space. The classification process results in a probability distribution for each document across topics such that the scores add to one and describe how prevalent that topic is in that response. The paper reports each topic's average prevalence across responses and the share of responses in which it is the dominant topic. Without regard to a topic model, the paper also reports the average length, grade level of language used, and positive and negative sentiment of responses.

To evaluate the reliability of automated scoring, for a sample of responses to the class-

room management question, we compare machine and human classification. Machine-human agreement is as high as human-human agreement, suggesting that automated scoring performs as well as human scoring here.

To evaluate the validity of the topic model technique, we derive and test hypotheses for within-applicant, across question-topic correlations in substantively connected topics using the two most commonly answered questions. Among applicants who answered both questions, almost all hypothesized positive relationships are positively and significantly related, suggesting that the topic modeling enables interpretable, reliable scores.[4]

Finally, we examine the measurement properties and practical implications of the resulting topic scores along two dimensions. First, we assess whether topic emphasis is associated with applicant characteristics and whether stylized topic-based selection rules could generate extreme demographic selection ratios, providing a diagnostic for potential disparate-selection concerns. Second, we examine across-response dispersion in topic mixtures within each question, measuring whether applicants answer a given question in substantively different ways or instead provide relatively similar topical profiles. This helps identify which questions generate a stronger candidate-differentiating signal and which appear more uniform or potentially less informative for screening. Together, these analyses show how the proposed measures can be used not only to describe applicant responses, but also to evaluate the practical risks and usefulness of open-ended screening questions.

Related work uses text analysis to study open-ended writing in teacher preparation and educator hiring contexts. Bartanen et al. (2025) analyze pre-service teachers' responses to a motivation-for-teaching prompt and relate the resulting topics to teacher preparation and early-career outcomes. The most closely-related paper is Liu et al. (2025), that studies required short essays in educator applications to one large, urban school district. The analysis focuses on three essay questions asked of all applicants and uses an existing method to relate essay features and themes to district ratings, hiring, retention, and teacher value-added. In contrast, we study a broader bank of questions and develop an inductive workflow to discover what each question elicits. In future work, we will relate these measures to hiring decisions and post-hire effectiveness and retention across multiple employers.

[4]Based on the meaning of the topics defined for these two questions, members of the research team independently identified which topics from the first question they expected to positively correlate with which topics from the second question. We discuss this validation process in detail in subsection 4.3.

2 Setting and Data

Our analytic dataset comes from the Nimble hiring platform.[5] The raw Nimble data include 693 unique districts, 2,127 district-years, and 756,181 applications. We restrict the data to hiring entities with substantial platform use and to district-years with at least one observed question response.[6] The final analytic sample includes 187 unique districts observed across 951 district-years, 72,824 applications, and 60,487 unique applicant-question responses. Within this sample, the modal number of questions observed in a district-year is 2, the mean is 2.22, the standard deviation is 1.29, and the maximum is 6.

For each question in Nimble's question bank, Table 1 presents a shorthand label, the full question prompt, average linguistic features of applicant responses, the number of applicant responses, and the number and share of district-years in which each question is asked. Across district-years, the most frequently asked question is, "How do you build a classroom culture in which all students feel valued?" which we reference using the shorthand label *students feel valued*. This question appears in 28.3% of district-years. The next most common questions ask candidates to describe the qualities that make them stand out (*candidate qualities*, 22.7%) and their classroom management style (*classroom management*, 20.8%).

Table 1: Bank of Questions and Average Response Characteristics

***Abbreviation* and Full Text**	**Word Length**	**Grade Level**	**Pos. Sent.**	**Neg. Sent.**	n	**District-years Num.**	**%**
classroom activities: If I were to walk into your classroom what activities would I see you and the students doing?	67.4 (0.88)	11.1 (0.05)	0.13 (0.00)	-0.02 (0.00)	6,296	74	12.2
students feel valued: How do you build a classroom culture in which all students feel valued?	78.8 (0.95)	9.9 (0.05)	0.23 (0.00)	-0.01 (0.00)	5,896	172	28.3
learning philosophy: What is your philosophy on learning?	71.3 (1.65)	9.5 (0.07)	0.22 (0.00)	-0.01 (0.00)	4,770	91	15.0

Continued on next page

[5]For simplicity, we use "district" to refer to the hiring entity that selects the question set. Some "districts" are charter management organizations or individual schools operating within such organizations.

[6]We define substantial platform use as being a paying client or, for districts whose access is paid for by a state department of education, having more than 100 applications over the sample period. This restriction leaves 291 unique districts, 1,524 district-years, and 745,058 applications. Restricting further to district-years with at least one observed question response yields the final analytic sample.

Table 1: Bank of Questions and Comparison Across Key Metrics (continued).

***Abbreviation* and Full Text**	**Word Length**	**Grade Level**	**Pos. Sent.**	**Neg. Sent.**	n	**District-years**	
						Num.	**%**
classroom management: Describe your classroom management style. Why have you found this style to be effective?	76.6 (1.05)	9.8 (0.06)	0.23 (0.00)	-0.01 (0.00)	4,629	126	20.8
student progress tracking: How do you track student progress in your classroom?	59.8 (1.34)	9.9 (0.09)	0.11 (0.00)	-0.01 (0.00)	1,910	76	12.5
philosophy learning growth: Describe your philosophy of how children learn and grow.	114.1 (3.74)	10.3 (0.08)	0.25 (0.00)	-0.00 (0.00)	1,622	74	12.2
candidate qualities: What qualities do you bring to this position which would make you the best candidate for the job?	101.5 (2.94)	10.9 (0.16)	0.24 (0.00)	-0.00 (0.00)	1,465	138	22.7
family communication: What do you feel is the most effective way to communicate with families? Describe how you have used this/these technique(s).	98.5 (2.00)	10.2 (0.09)	0.27 (0.00)	-0.00 (0.00)	1,342	110	18.1
active engagement: What does active engagement look like in your classroom?	50.0 (1.65)	11.1 (0.13)	0.11 (0.00)	-0.03 (0.00)	1,164	43	7.1
accommodate learning styles: How do you accommodate different learning styles in your classroom?	64.3 (2.29)	10.9 (0.22)	0.15 (0.01)	-0.01 (0.00)	817	89	14.7
difficult feedback: What do you do to improve? Give an example of a time when you received difficult feedback from a peer/a student/a teacher/a supervisor. How did you respond in the moment? How did it affect your thinking? Your actions?	118.0 (3.66)	9.1 (0.12)	0.16 (0.00)	-0.01 (0.00)	664	56	9.2
feedback sources: To whom can you look for valuable feedback? How did you respond to that feedback?	55.5 (3.33)	8.4 (0.17)	0.20 (0.01)	-0.01 (0.00)	434	46	7.6

Note: Columns report average linguistic and sentiment characteristics for responses to each question. *Word Length* and *Grade Level* capture textual complexity; *Grade Level* is an approximate readability grade-equivalent measure of surface-level text complexity and should not be interpreted as applicant education level or response quality. *Positive* and *Negative Sentiment* indicate average sentiment polarity. n denotes the number of valid responses for each question used in the analysis. *District-years* reports the number and percent of district-years in the sample that asked the specific question.

Across questions, average characteristics are based on response sample sizes ranging from 434 to 6,296. Response length varies substantially across questions, from an average of 50 words for *active engagement* to 118 words for *difficult feedback*. Questions focused on growth, improvement, or broader learning philosophy tend to elicit longer responses. In contrast, questions focused on concrete classroom practices, such as active engagement and student progress tracking, tend to elicit shorter responses.

To measure the linguistic complexity of responses, we compute a readability grade level for each response[7] and then average these scores within each question. Average grade-level scores range from 8.4 to 11.1. Higher scores indicate that responses tend to use more complex surface-level language, such as longer words or sentence structures, rather than necessarily better, more accurate, or more job-relevant content. This distinction is important: readability measures capture features of written expression, not the substantive quality of an applicant's answer. Some questions place greater demands on written fluency and linguistic complexity than others, which may affect how these questions function as screening instruments. Responses to *classroom activities* and *active engagement* have the highest average grade-level scores (11.1), whereas *feedback sources* responses have the lowest (8.4).

We compute sentiment scores for responses to each interview question, providing a diagnostic for whether different questions elicit systematically different emotional tones in applicants' written narratives[8]. Question prompt framing may lead applicants to emphasize successes and enthusiasm, yielding more positive valence, or challenges and conflict, yielding more negative valence. Sentiment scores therefore facilitate the identification of response-style differences across questions and potential construct-irrelevant variation. If some questions consistently elicit a more positive or negative tone, this may reflect question-induced framing or strategic self-presentation rather than differences in job-relevant content, with implications for comparability and fairness. Responses are generally neutral to posi-

[7]Readability grade levels are approximate grade-equivalent measures of surface-level text complexity (Flesch, 1948; Kincaid et al., 1975; Chall and Dale, 1995). For example, a score near 8 suggests language with sentence and word features typical of text readable around the eighth-grade level, while a score near 11 suggests text closer to the eleventh-grade level. These scores should not be interpreted as measuring the applicant's education level, the quality of the response, or the job relevance of the content.

[8]Sentiment scores are derived from TextBlob polarity scores, which range from -1 to 1 for each response, with values closer to zero indicating more neutral language, larger positive values indicate more positive affective tone, and more negative values indicate more negative affective tone. We decompose each response's polarity score into two components: Positive Sentiment, equal to the polarity score when positive and zero otherwise (ranging from 0 to 1), and Negative Sentiment, equal to the polarity score when negative and zero otherwise (ranging from -1 to 0). Reported averages therefore reflect the average positive or negative affective content of responses to a given question. For example, language emphasizing enthusiasm, collaboration, growth, and successful relationships receives a more positive score, whereas language emphasizing conflict, difficulty, frustration, or behavioral problems receives a more negative score. The measure should be interpreted as capturing surface-level affective tone rather than the substantive quality or job relevance of a response.

tive in tone: average positive sentiment ranges from 0.107 to 0.270, while average negative sentiment remains close to zero, ranging from -0.028 to -0.002. The *family communication* question generates the most positive sentiment (0.270), followed by *philosophy learning growth* (0.246) and *candidate qualities* (0.244). By contrast, *active engagement* produces both the lowest positive sentiment score (0.107) and the most negative average sentiment score (-0.028), suggesting that discussions of student engagement elicit more balanced or problem-focused language. Response length, linguistic complexity, and affective tone differ across applicants and their distributions differ across questions.

3 What topics do applicants discuss in their answers to each question?

For each question, the analysis involves two main tasks: topic discovery and document classification. Topic discovery defines a parsimonious and substantively meaningful set of topics that describes the range of applicant responses to a question. We adapt ITA by combining researcher judgment with LLMs and topic-modeling tools. This hybrid approach is a methodological contribution because it scales key elements of ITA while preserving transparency and interpretability. It also offers interpretability and auditablity advantages over fully unsupervised topic-discovery methods, such as LDA or direct embedding-based clustering of raw responses, and over prompt-only LLM-assisted thematic analysis. The output of topic discovery is a set of topics, each accompanied by a written description.

The second task is document classification. After defining a set of topics for a question, we use LLM-based classifiers to score each response with respect to the topics. The topic descriptions from the discovery serve as classification instructions. The resulting document-topic scores convert any open-text response into quantitative measures that can be used in subsequent analysis. We validate the automated scoring by comparing LLM-generated classifications with human ratings and use these comparisons to inform topic refinement.

This section describes the method in detail for the most commonly answered question: *classroom management*. We apply the same workflow to the remaining questions and report key outputs for each in Appendix A.8. Presenting one case in detail allows readers to see how the method operates while keeping the output from the full set of questions available for audit in the appendix.

3.1 Background on Inductive Thematic Analysis and LLM use

Thematic analysis is a foundational qualitative method for identifying and interpreting patterns, or themes, in textual data. Braun and Clarke (2006) provided an influential structured framework for this approach. Before their contribution, researchers used a variety of qualitative coding strategies, but there was less consensus on what thematic analysis entailed (Boyatzis, 1998). Braun and Clarke (2006) outline a six-phase process: 1) familiarization with the data, 2) generating initial codes, 3) searching for themes, 4) reviewing themes, 5) defining and naming themes, and 6) producing the report. The process is iterative rather than linear, allowing researchers to refine themes as they engage with the data.

Researchers often utilize computer-assisted qualitative data analysis software, such as NVivo, MAXQDA, and ATLAS.ti, to search for and retrieve text, and they also use technology to suggest patterns in large text corpora (e.g., LDA). These tools can make qualitative analysis more scalable and improve documentation and replicability (Woods et al., 2016). Even with the use of technology tools, however, ITA remains labor-intensive because it typically requires human researchers to code text, iteratively refine codes, and derive themes grounded in the data.

Recent advances in LLMs have generated interest in using these models to assist with, or partially automate, qualitative coding and theme development (e.g., Zhang et al. 2023, De Paoli 2024, Wang et al. 2025). LLMs can summarize, interpret, and classify text at scale, making them potentially useful for early coding and theme-generation tasks in large datasets (Jalali and Akhavan, 2024). Since the public release of ChatGPT in late 2022, a growing body of work has evaluated LLMs as qualitative research assistants (Long, 2024).

Academic interest in LLM-assisted qualitative coding accelerated around 2023. Early studies have demonstrated feasibility. De Paoli (2024) used GPT-3.5 for ITA on human-coded interview transcripts from prior studies and found that AI-generated codes captured most human-identified themes with a "good degree of validity" after prompt engineering. Zhang et al. (2023) evaluated API-based calls to OpenAI's GPT models for qualitative coding and found that the tool can refine the coding process while improving transparency, credibility, and accessibility relative to traditional qualitative analysis software. Relatedly, Wang et al. (2025) introduced LLM-Assisted Thematic Analysis (LATA), a prompt-structured application of Braun and Clarke's six-phase framework using GPT-4 and Gemini, benchmarked against human coding. Their study demonstrates that LLMs can be prompted to perform the stages of thematic analysis and produce theme summaries comparable to human-generated analysis in a small interview corpus. Our approach builds on this literature but differs in objective and design: rather than asking an LLM to conduct thematic analysis end-to-end, we combine LLM-generated codes with topic-model-based clustering, researcher-defined topic

refinement, and LLM-based probabilistic classification to produce auditable response-level measures for quantitative analysis. Chen (2024) build on Cascio et al. (2019), who recommend a team-based approach to inductive coding and assessment of agreement across coders and LLMs. They introduce a theory-informed computational method for evaluating outputs from inductive coding.[9]

3.2 Hybrid LLM-–Topic-Model Inductive Thematic Analysis (HLTM-ITA) of Response Text

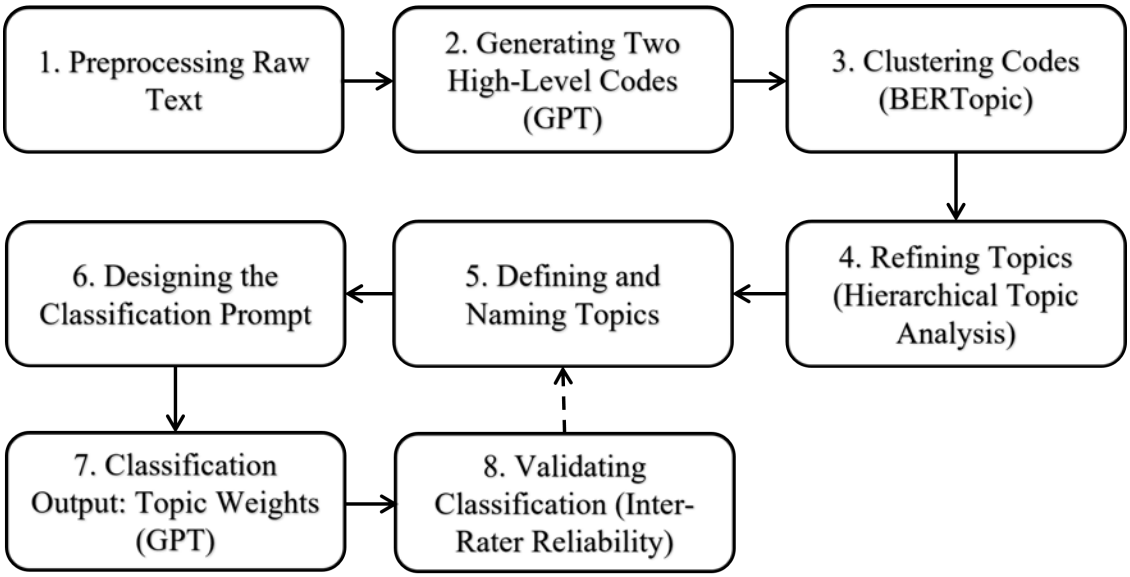
In this section we describe our Hybrid LLM–Topic-Model Inductive Thematic Analysis, a topic-discovery and classification workflow that adapts and scales phases 1–5 of ITA (Braun and Clarke, 2006) while preserving traceability from final topics back to the original response text. Figure 1 summarizes the workflow. Steps 1–5 conduct topic discovery and parallel the first five phases of ITA. Steps 6–8 translate the resulting topics into quantitative response-level measures: Step 6 converts the topic definitions into a classification prompt, Step 7 applies that prompt to assign topic weights to each response, and Step 8 validates the resulting classifications using inter-rater reliability. When validation reveals ambiguity or low agreement, results from Step 8 feed back into Step 5, allowing researchers to revise the topic definitions and then rerun the classification step. This section describes each step using the *classroom management* question as the running example. For those wanting more detail on the process and the exact prompt language used, Appendix A gives details. We then apply the same process to all questions listed in Table 1, with question-specific outputs reported in subsection A.8.

HLTM shares some commonalities with prompt-driven LLM-assisted thematic analysis workflows. In particular, both use LLMs to support early coding of raw response text (e.g., Zhang et al. 2023, De Paoli 2024, Wang et al. 2025). The key difference is that HLTM does not rely on prompts alone to conduct the full thematic analysis or to define the final topics. Instead, we use the LLM to generate traceable codes from each response, use BERTopic (Grootendorst, 2022) to group those codes into embedding-based clusters, and then use researcher judgment to define the final topics used for classification. This intermediate clustering step improves scalability while preserving an auditable path from final topics back to clusters, codes, representative quotes, and original responses. The final output also differs.

[9]Applications are expanding across disciplines, and a full review is beyond the scope of this paper. For example, Long (2024) used GPT-4 to code classroom dialogue and found high agreement with experts and substantial time savings. Jalali and Akhavan (2024) applied ChatGPT to public health interviews and found similar themes but weaker contextual interpretation, suggesting that AI tools may be most useful for early coding and theme generation.

Rather than producing only a qualitative thematic report, HLTM produces response-level topic scores that can be validated against human ratings and used in quantitative analysis.

Figure 1: End-to-end semantic analysis pipeline



Note: The figure summarizes the HLTM workflow. Steps 1–5 conduct topic discovery and parallel phases 1–5 of Braun and Clarke (2006). Steps 6–8 convert the discovered topics into response-level quantitative measures. The dashed arrow from Step 8 to Step 5 indicates that validation results can motivate revisions to the topic definitions before rerunning classification.

In **Step 1**, we preprocess response documents to prepare them for analysis and to remove responses that are unlikely to contain meaningful information. This step has two components. First, we apply deterministic cleaning rules that remove very short, blank, duplicate, or near-duplicate responses.[10] Second, we identify responses that appear to be data-entry errors or intentionally uninformative answers. To do so, we use three independently worded GPT prompts to classify each response as responsive or nonresponsive. A response is removed only if all three prompts classify it as nonresponsive. This procedure removed 170 responses for *classroom management* and 252 responses for *students feel valued*. Appendix A.2 reports the prompts and additional details.[11] Although nonresponsive answers may be informative in later models of hiring or post-hire outcomes, they add noise to topic discovery and are therefore excluded from the thematic analysis.

[10] Appendix A.1 details preprocessing, using the *classroom management* question as an example.

[11] The three OpenAI API calls used prompts with different levels of detail: long, medium, and short. The most common errors appeared to occur when applicants inadvertently entered a response intended for another field, producing a valid answer to some question but not to the question under analysis. Some applicants also entered intentionally uninformative text to bypass required questions.

Table 2: Illustrative Applicant Responses and Associated Codes

Input: Applicant's Response	**Outputs:**			
	Quote	**Code Name**	**Code Description**	**Code ID**
My style is highly motivating and creative. I love truth and respect—as long as no one is disrespected, all is well. I do not tolerate negative attitudes, ever. I understand that children might come from difficult households. So, I will communicate with each student according to their personality, all within the context and guidelines of class curriculum and school policy.	"my style is highly motivating and creative (...) I do not tolerate negative attitudes"	Motivational Environment	Emphasizes creating a motivating and creative atmosphere that fosters respect and positivity among students. This approach aims to uplift students.	841
	"I will communicate to with each student according to their personality"	Personalized Communication	Highlights understanding student's personality to tailor communication effectively, ensuring that interactions align with curriculum and school policies.	842
My classroom management style is respectful, caring, structured. The students are all aware of the classroom rules and procedures, they are also aware of the consequences because I discuss them the first few weeks of class and I set a clear expectations early on in the school year.	"The students are all aware of the classroom rules and procedures."	Respectful Environment	This theme emphasizes the importance of creating a caring and respectful atmosphere in the classroom, where students feel valued and understood.	681
	"I set clear expectations early on in the school year"	Clear Expectations	Establishing clear expectations from the beginning helps students understand their responsibilities and the structure of the classroom, promoting a positive learning environment.	682
I am fair and consistent. In Physical Education, I want all students to experience some kind of success and have fun at the same time. I find peer teaching works very well.	"I am fair and consistent."	Fairness and Consistency	This theme emphasizes the importance of treating all students equally and maintaining a stable environment for learning.	11582511
	"I find peer teaching works very well."	Peer Teaching Benefits	This theme focuses on the effectiveness of collaborative learning strategies, where students learn from each other, enhancing their understanding and engagement.	11582512

Continued on next page

Table 2: Illustrative Applicant Responses and Associated Codes (continued)

Applicant's Response	Quote	Code Name	Code Description	Code ID
My classroom management style is my strongest skill. I believe in clear expectations, if expectations are not met then at that point logical consequences are applied.	"I believe in clear expectations."	Clear Expectations	Establishing clear expectations is fundamental to effective classroom management, ensuring students understand what is required of them.	76551
	"If expectations are not met then at that point logical consequences are applied."	Logical Consequences	Implementing logical consequences helps maintain order and encourages accountability among students.	76552

Step 2 generates initial codes. In this stage, an LLM processes each response independently and identifies the most prominent ideas in the text. Following ITA terminology, we refer to these ideas as codes. The model is prompted to extract up to two codes per response and, for each code, to provide a concise label of no more than three words, a brief description, and a representative quote from the original response.[12] Each code is assigned a unique identifier, allowing it to be traced back to the response from which it was generated. Appendix A.3 reports the full prompt. Each response is processed through a separate API call, so the model has no memory of other responses when coding a given document.

For the *classroom management* question ("Describe your classroom management style. Why have you found this style to be effective?"), this produced 9,025 codes from 4,629 applicant responses, or 1.95 codes per response. Shorter responses sometimes produced only one code. Of the 9,025 generated codes, 2,646 code names are unique. Table 2 provides examples of the output from Step 2.[13] For example, the table's first response is represented by two codes: "Motivational Environment" and "Personalized Communication." Each code includes a description, a quote drawn directly from the response, and a unique code ID.

In **Step 3**, initial codes are grouped into broader clusters.[14] We use BERTopic (Grootendorst, 2022) to cluster codes based on the code label, code description, and representative

[12]We tested prompts requesting one, two, or up to three codes per response. One code missed some substantive detail, while allowing up to three codes added little new information. The final topics were similar across these alternatives.

[13]Appendix A.4 reports additional information on the generated codes and their frequency distribution.

[14]In prompt-driven LLM-assisted thematic analysis workflows, researchers often prompt the LLM to search for, review, and define themes from the initial codes (Zhang et al., 2023; De Paoli, 2024; Wang et al., 2025). This approach is useful for exploratory qualitative analysis but leaves much of the code-to-theme grouping process inside the LLM-generated response. We instead introduce an intermediate, model-based clustering step.

quote generated in Step 2. BERTopic embeds the code-level text—using code label, description, and quotes as input—reduces dimensionality, clusters semantically similar codes, and generates topic representations for the resulting clusters. We tune the model separately for each question and select among specifications using semantic coherence and related diagnostics.[15] Applying BERTopic to the 9,025 classroom-management codes produced 35 embedding-based clusters. The OpenAI representation model was then used to generate labels and descriptions for these clusters.[16] Table 3 reports the 35 clusters and the number of codes assigned to each.

Table 3: BERTopic Code Clusters and Raw Code Counts for the *classroom management* Question

Cluster ID	**Cluster Name**	**Freq. of Codes**
0	Student Engagement and Active Participation	983
1	Positive Reinforcement in Education	788
2	Establishing Clear Classroom Expectations	651
3	Building Strong Relationships with Students	523
4	Mutual Respect in Classroom Environment	482
5	Authoritative Classroom Management Style	355
6	Collaborative Learning and Teamwork	282
7	Structured Classroom Environment for Learning	264
8	Student Empowerment and Ownership	207
9	Proactive Classroom Management Strategies	201
10	Student Accountability and Responsibility	193
11	Flexible Teaching and Adaptability	153

Continued on next page

[15] We explored several alternatives for converting text to quantitative data, including LDA, BERTopic applied directly to raw responses, and prompt-only LLM-assisted thematic analysis using API calls to various LLMs (e.g., Zhang et al., 2023; De Paoli, 2024; Wang et al., 2025). We judged that LDA and BERTopic applied directly to raw responses required many top-down impositions of researcher discretion in specifying the topical structure, while prompt-only LLM-assisted thematic analysis weakened traceability from final themes back to the original text because intermediate grouping decisions were harder to audit. A companion methodological report provides implementation details and comparisons across methods (Aguilar et al., 2025).

[16] Following the tuning strategy of Grootendorst (2022), we optimize BERTopic separately for each corpus. The grid varies two embedding back ends—the default `all-MiniLM-L6-v2` and OpenAI's `text-embedding-3-large`—together with four clustering hyperparameters: UMAP neighborhood size $n_{\text{neighbors}} \in \{5, 15, 50\}$, UMAP dimensionality $n_{\text{components}} \in \{5, 10, 20\}$, HDBSCAN minimum cluster size min_cluster_size $\in \{5, 15, 50, 150\}$, and HDBSCAN minimum samples min_samples $\in \{5, 15, 50, 150\}$. We assess the full $2 \times 3 \times 3 \times 4 \times 4 = 288$ grid using clustering diagnostics and the semantic coherence measure c_v. Because c_v most directly captures topical interpretability, we retain the configuration maximizing this metric. For the classroom-management corpus, the selected specification uses MiniLM embeddings with $n_{\text{neighbors}} = 50$, $n_{\text{components}} = 20$, min_cluster_size $= 50$, and min_samples $= 5$.

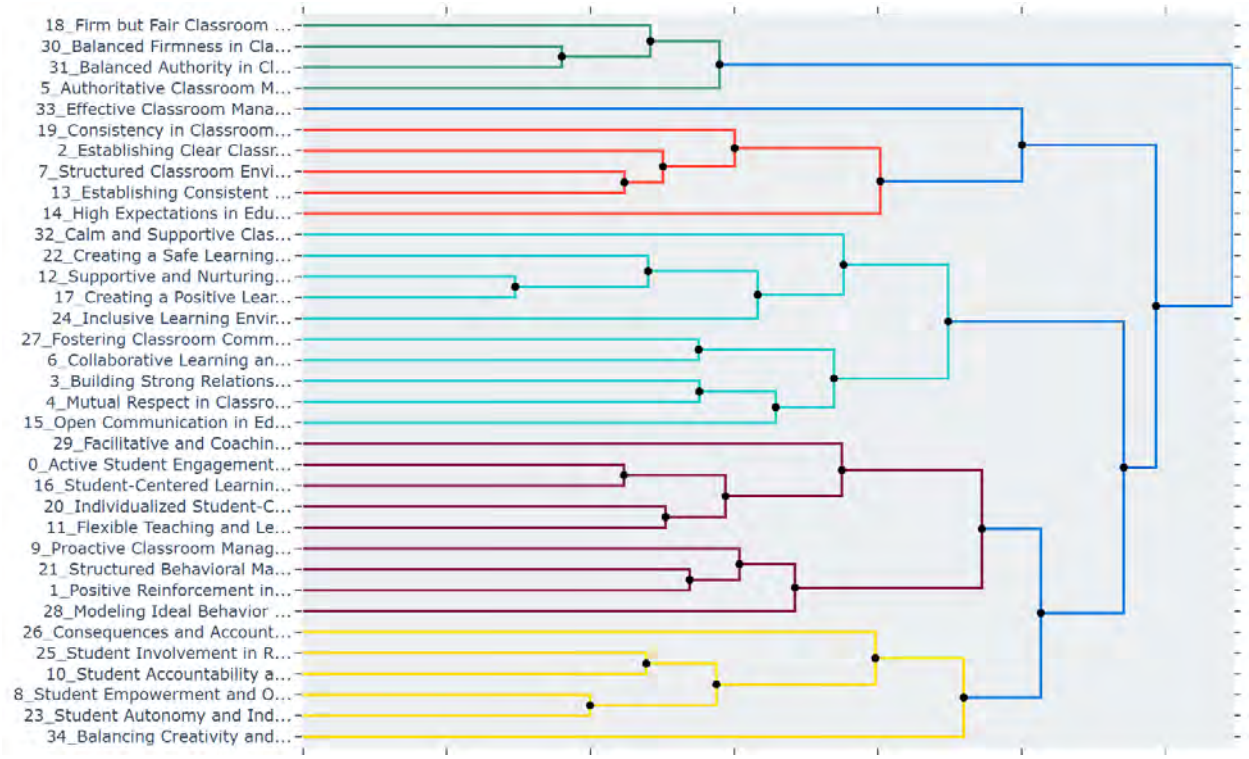
Table 3: BERTopic Code Clusters and Raw Code Counts for the *classroom management* Question (continued)

Cluster ID	Cluster Name	Freq. of Codes
12	Supportive and Nurturing Learning Environment	151
13	Establishing Consistent Classroom Routines	148
14	High Expectations in Education	136
15	Open Communication in Education	131
16	Student-Centered Learning and Approach	130
17	Creating a Positive Learning Environment	118
18	Firm but Fair Classroom Management	108
19	Consistency in Classroom Management Practices	96
20	Individualized Student-Centered Learning Approaches	91
21	Structured Behavioral Management Approaches	89
22	Creating a Safe Learning Environment	85
23	Student Autonomy and Independence	82
24	Inclusive Learning Environment for All	80
25	Student Involvement in Rule Creation	72
26	Clear Rules and Consequences Implementation	70
27	Fostering Classroom Community and Belonging	67
28	Modeling Ideal Behavior in Classrooms	61
29	Facilitative and Coaching Teaching Approach	43
30	Balanced Firmness in Classroom Management	24
31	Balanced Authority in Teaching	15
32	Calm and Supportive Classroom Environment	13
33	Effective Classroom Management Planning	7
34	Balancing Creativity and Structure	6

Note: Cluster IDs correspond to BERTopic-generated clusters. Cluster names are human-readable labels assigned by the OpenAI representation model to summarize the content of each cluster. "Freq. of Codes" reports the number of individual codes assigned to each cluster. The BERTopic model was fit on 9,025 codes in total. Of these, 2,120 were assigned to the outlier cluster (-1), which is not shown in the table. Therefore, the frequencies reported here sum to 6,905.

These outputs support **Step 4**, in which researchers review and refine candidate topic groupings. Figure 2 shows the hierarchical structure of the labeled 35 BERTopic clusters for the classroom management question. The hierarchy helps researchers identify sets of clusters that may be combined into more parsimonious topics for subsequent analysis. Because the tree emphasizes hierarchical merging rather than absolute semantic distance, we also use two-dimensional cluster visualizations to assess how close clusters are in the embedding space.

Figure 2: *classroom management* Question: Hierarchical Tree Representation of the BERTopics Structure



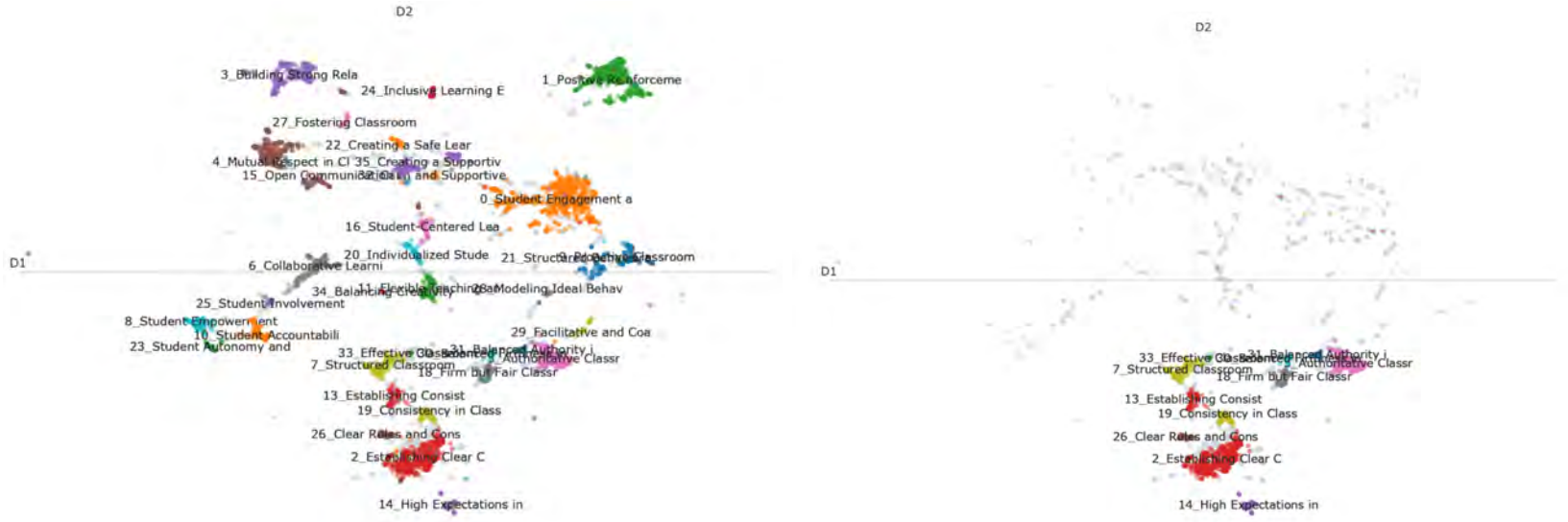
Note: The figure visualizes the hierarchical structure of the 35 clusters reported in Table 3. Clusters are ordered by thematic similarity, with finer groupings on the left and more aggregated groupings on the right. Researchers use this representation to identify possible aggregations of code clusters into final topics.

Figure 3 shows two-dimensional representations of the cluster structure at hierarchy level 0, the least aggregated level, and level 8, one of the most aggregated levels. We also inspect the underlying codes, descriptions, and quotes assigned to each cluster to verify that the clusters are substantively coherent. Appendix Table A.4 provides an example for the four largest clusters, where the overwhelming majority of codes are very closely related to the broader theme.

We use the hierarchical tree in Figure 2, the two-dimensional representations in Figure 3, and direct inspection of the underlying codes to gain insight into how clusters relate to one another and to guide topic refinement.[17] The hierarchical tree provides the primary structure for identifying candidate merges: clusters that merge at shorter distances in the dendrogram, or that appear within the same nearby branch, are treated as candidates for combination. BERTopic orders clusters according to their estimated similarity, and colored branches in Figure 2 provide a visual cue for broad groupings, but we treat them as diagnostic aids rather

[17]BERTopic also produces a text-based version of the same hierarchy that reports higher level node labels for potential merged topics. We used these labels as interpretive aids; final topic definitions were based on the visual hierarchy, the two-dimensional representations, and inspection of the underlying text.

Figure 3: Two-Dimensional Representations of Documents and BERTopic Clusters for *classroom management* Question



(a) All Clusters

(b) Topic 1 Clusters

Note: The figure visualizes the semantic structure of the *classroom-management* code clusters. Points represent codes, and labels correspond to BERTopic clusters. Panel (a) shows all clusters in their least aggregated structure, while Panel (b) shows only the clusters corresponding to topic 1: Authoritative Approach for Structure and Accountability. Gray points indicate codes assigned to the residual group. Researchers use these plots, together with the hierarchy in Figure 2, to evaluate possible aggregations of clusters into final topics.

than mechanical decision rules. The two-dimensional document-topic visualization provides a complementary check by showing whether candidate clusters occupy nearby regions in the reduced embedding space. The procedure proceeds iteratively. We begin with adjacent clusters in the hierarchy, examine their positions in the two-dimensional plot, and review the underlying cluster labels, keywords, descriptions, representative code-level documents, and code examples. If clusters are close both visually and conceptually, we combine them into a candidate topic. If they are semantically distant or capture distinct conceptual content, we leave them separate or assign them to different emerging topics. We repeat this process until we arrive at a smaller set of coherent and interpretable topics.

For example, in the *classroom management* question, the final topic *Authoritative Approach for Structure and Accountability* was formed by combining clusters that emphasized authority, firmness, structure, expectations, routines, consistency, and consequences. In Figure 2, clusters 18, 30, 31, and 5 appear together in the upper branch, shown in green[18]: *Firm but Fair Classroom Management*, *Balanced Firmness in Classroom Management*, *Balanced Authority in Teaching*, and *Authoritative Classroom Management Style*. These clusters were treated as a first candidate subgroup because they merge within the same branch and because their labels and underlying codes describe a common idea of balanced authority: being firm, fair, consistent, and authoritative without being punitive.

A second nearby branch in Figure 2 contains clusters 33, 19, 2, 7, 13, and 14[19]: *Effective Classroom Management Planning*, *Consistency in Classroom Management Practices*, *Establishing Clear Classroom Expectations*, *Structured Classroom Environment for Learning*, *Establishing Consistent Classroom Routines*, and *High Expectations in Education*. We treated these clusters as a second candidate subgroup because they emphasize predictable routines, clear expectations, planning, consistency, and high standards. Although the two subgroups are not identical, direct inspection of their code labels, descriptions, and representative code-level text showed that they describe complementary parts of the same broader classroom-management orientation: an authoritative approach implemented through clear structure and accountability. We therefore treated them as components of the same emerging topic.

The two-dimensional visualization in Figure 3 provides additional information on this decision. When the candidate clusters are isolated in the two-dimensional document-topic plot, the authority-and-firmness clusters and the structure-and-expectations clusters appear in nearby and partially overlapping regions of the reduced embedding space. This visual

[18]The text-based version of the hierarchical tree identifies this parent branch as *Authoritative Classroom Management Style*.

[19]The same text-based version of the hierarchical tree identifies this nearby branch as *Establishing Clear Classroom Expectations*.

pattern supports, but does not mechanically determine, the decision to collapse them into a single topic rather than treat them as unrelated dimensions. Cluster 26, *Clear Rules and Consequences Implementation*, illustrates why substantive inspection remains necessary. In the hierarchical tree, cluster 26 appears in the lower branch, shown in yellow, separate from the two main branches described above. However, its underlying codes and representative code-level documents emphasize rules, consequences, accountability, and behavioral expectations, and the two-dimensional plot places it near the same region as the structure-and-expectations clusters. We therefore included cluster 26 in the same final topic. This example illustrates the full decision process: the hierarchy identifies candidate merges, the two-dimensional plot checks whether candidate clusters occupy nearby regions in the reduced representation, and researcher inspection of the underlying text determines whether the clusters belong to the same final topic.

In **Step 5**, researchers select the final topic scheme for each question. This choice is informed by the generated codes, BERTopic clusters, hierarchical tree, two-dimensional visualizations, and direct inspection of representative quotes. For each question, two or more researchers independently reviewed these materials and proposed a topic structure. The research team then discussed these proposals and agreed on a final set of topics.

In **Step 6**, researchers write topic descriptions to be concise, substantively distinct, and useful for subsequent classification. Topic descriptions were informed by the researchers' review of the codes, clusters, and representative quotes. We avoided repeating concepts or keywords across topics when possible. When topic descriptions were too similar or captured overlapping concepts, we merged them to improve classification consistency.

For the *classroom management* question, this process initially produced a six-topic scheme. Table 4 reports the six topics, the BERTopic clusters used to construct each topic, and the most frequent codes associated with those clusters. The method therefore produces a transparent, traceable record of interpretation: each final topic can be linked to its component clusters, each cluster can be linked to its underlying codes and quotes, and each code can be linked to the original applicant responses.

Table 4: Initial Six-Topic Scheme for the *classroom management* Question

Topic	**Cluster IDs**	**Associated Codes (Frequency)**
Authoritative Approach for Structure and Accountability	18, 30, 31, 5, 33, 19, 2, 7, 13, 14, 26	Authoritative Approach (309) Clear Expectations (301) Structured Environment (123) High Expectations (101) Organized Environment (33) Expectation Setting (25) Rules and Expectations (23) Expectations Setting (19) Behavior Expectations (18) Routine and Structure (17) Rule Establishment (17) Firm Yet Fair (16) Behavioral Expectations (12) Clear Communication (12) High Standards (12) **Total Frequency: 1,874**
Student Engagement and Empowerment	25, 10, 8, 23, 34, 0	Student Engagement (225) Student Empowerment (142) Engagement Focus (98) Engagement Strategies (88) Student Involvement (61) Active Engagement (48) Student Autonomy (32) Engaging Environment (29) Interactive Learning (24) Fun Learning Environment (22) Accountability Focus (21) Engagement Techniques (21) Engaging Lessons (21) Student Accountability (20) Student Responsibility (19) **Total Frequency: 1,543**
Student-Centered Adaptive Teaching Approaches	29, 16, 20, 11	Student-Centered Approach (67) Student-Centered Learning (30) Individualized Approach (17) Adaptability in Teaching (14) Flexibility in Teaching (13) Flexible Approach (13) Individualized Support (12) Individual Attention (11) Facilitative Approach (10) Individualized Learning (9) Coaching Approach (8) Adaptive Strategies (7) Adaptive Teaching (6) Flexibility in Approach (6) Personalized Learning (6) **Total Frequency: 417**
Building Strong Relationships	32, 22, 12, 17, 24, 27, 6, 3, 4, 15	Relationship Building (167) Respectful Environment (139) Supportive Environment (127) Mutual Respect (76) Collaborative Learning (73) Positive Environment (72) Positive Relationships (67) Building Relationships (49) Safe Learning Environment (40) Collaborative Environment (35) Community Building (35) Inclusive Environment (32) Open Communication (29) Building Rapport (23) Positive Learning Environment (22) **Total Frequency: 1.932**

Continued on next page

Table 4: Initial Six-Topic Scheme for the *classroom management* Question (continued)

Topic	**Cluster IDs**	**Associated Codes (Frequency)**
Positive Reinforcement Strategies and Systems	9, 21, 1	Positive Reinforcement (512) Proactive Approach (70) Proactive Engagement (55) Proactive Environment (28) Positive Discipline (23) Reward System (23) Reward Systems (11) Intrinsic Motivation (9) Positive Behavior Support (8) Behavior Management (7) Behavior Correction (6) Behavioral Strategies (5) Encouragement Focus (5) Motivation Techniques (5) Positive Behavior Focus (5) **Total Frequency: 1,078**
Modeling Positive Behavior by Example	28	Modeling Behavior (29) Behavior Modeling (12) Role Model Approach (2) Behavioral Principles (1) Behaviorist Approach (1) Business Model Approach (1) Character Building (1) Character Development (1) Modeling Techniques (1) Modeling and Demonstration (1) Modeling and Reinforcement (1) Modeling and Relationships (1) Positive Behavior Modeling (1) Positive Role Model (1) Procedures and Modeling (1) **Total Frequency: 61**

Note: Cluster IDs correspond to the BERTopic clusters reported in Table 3. Associated Codes (Frequency) lists the 15 most frequent codes associated with each topic through the underlying clusters.

Step 7 uses the topic scheme from Step 5 to score applicant responses at scale. The topics and their descriptions become the classification rubric, and each response is assigned a set of topic weights. This step is fundamentally a scoring task. In settings where theory and domain expertise already provide a well-specified rubric, researchers often proceed directly to this type of classification without first conducting inductive analysis. Our setting differs because the relevant dimensions of information in applicants' written responses are not sufficiently specified ex ante. We therefore use inductive analysis to learn those dimensions from the text before applying them systematically. By contrast, Rubin and Grissom (2026) use LLMs in a closely related way to implement rating tasks in settings where experts can design scoring rubrics in advance. Their work is most closely related to our Step 6: once a coding scheme exists, LLMs can apply it consistently and at scale, provided that reliability, uncertainty, and validation are carefully assessed.

We prompt the LLM to read each response and assign probability weights to the identified topics based on their prevalence in the text. The weights are constrained to sum to one. When multiple topics are present, the prompt instructs the model to allocate greater weight to topics that are more prominent in the response. To improve consistency, the prompt

includes a brief author-coded example. We use OpenAI GPT models through the API to generate these topic-alignment scores. Appendix A.6 provides additional details and an example prompt. The output is the original response-level dataset augmented with one score for each topic.

In **Step 8**, we assess and report inter-rater agreement for the *classroom management* question scoring by comparing GPT-generated topic scores with human scores on a random sample of 200 responses. Two authors, denoted H1 and H2, followed the same instructions used in the GPT prompt. Because there is no objective ground truth for the topic composition of a response, human-human agreement provides the relevant benchmark. If LLM-human agreement is much lower than human-human agreement, the classification task may require better prompts, a different model, or clearer instructions. If LLM-human agreement is comparable to human-human agreement, ambiguity in the topic framework itself may be the binding constraint (i.e., topic definitions).

We find that human-machine agreement is quite similar to human-human agreement. Specifically, we compare agreement for three rater pairs: GPT with H1, GPT with H2, and H1 with H2. In Table 5, we report two rank-based agreement measures: *per-topic agreement*, measured by Cohen's κ; and *overall rank-order agreement*, measured by Spearman's ρ.[20] Cohen's κ is computed separately for each topic using raters' rank assignments and captures agreement on the rank assigned to that topic beyond what would be expected by chance. Spearman's rank correlation ρ summarizes overall agreement in the ordering of the six topics across raters. Higher values of both Cohen's κ and Spearman's ρ indicate greater agreement. Taken together, these measures assess whether GPT and human raters assign similar topic ranks, both for individual topics and for the overall ordering of topics within responses.

Table 5: Inter-Rater Reliability in Rank-Based Six-Topic Classification

Raters:	**GPT&H1**	**GPT&H2**	**H1&H2**
Per-Topic Cohen κ (Ranks)			
Authoritative Approach for Structure & Accountability	0.736	0.812	0.710
Student Engagement and Empowerment	0.643	0.528	0.506
Student-Centered Adaptive Teaching Approaches	0.536	0.447	0.335
Building Strong Relationships	0.558	0.681	0.502

Continued on next page

[20] Appendix A.7 reports additional agreement measures computed on the same validation sample. These include Kendall's τ for rank-order agreement and several rank-implied distributional measures constructed by mapping topic ranks into normalized topic-weight distributions: cosine similarity, Pearson's r, the square root of Jensen–Shannon divergence, $\sqrt{\text{JSD}}$, Hellinger distance, Wasserstein$_1$, and Brier score.

Table 5: Inter-Rater Reliability in Rank-Based Six-Topic Classification (continued)

Raters:	**GPT&H1**	**GPT&H2**	**H1&H2**
Positive Reinforcement Strategies and Systems	0.725	0.716	0.781
Modeling Positive Behavior by Example	0.345	0.506	0.643
Overall Rank-Order Agreement			
Spearman Rank Correlation ρ	0.700	0.730	0.668

Note: GPT = GPT-4o-mini. Human raters are denoted as H1 and H2, where H1 is the principal investigator and H2 is a PhD student. Reliability is computed on the same random sample of 200 responses. Per-topic Cohen's κ is computed from raters' rank assignments for each topic. Overall Spearman rank correlation ρ summarizes agreement in the ordering of topic scores across the six-topic classification. Higher values of both Cohen's κ and Spearman ρ indicate greater agreement.

We report reliability for both the initial six-topic framework and a refined four-topic framework that collapses conceptually overlapping topics (Tables 5 and 6). Because the final classification model uses the four-topic structure, comparing the two frameworks allows us to assess whether conceptual consolidation improves agreement.

The main question is whether human and machine agree on coding at a rate similar to how two human raters agree. Table 5 shows that human–human agreement (H1&H2) is broadly comparable to each human rater's agreement with GPT. This suggests that the LLM performs similarly to human raters in applying the six-topic instructions. Spearman rank correlations range from 0.67 to 0.73 across the three rater pairs, indicating moderate-to-strong but imperfect agreement in the ordering of topics. Relative to the human-human benchmark, GPT–human rank correlations are 0.700 and 0.730, compared with 0.668 for H1&H2. Agreement is stronger for some topics than others. Using the conventional benchmarks from Landis and Koch (1977), values of Cohen's κ between 0.61 and 0.80 are often described as "substantial" agreement. Five of the six topics have at least one rater-pair comparison in this range, while *Student-Centered Adaptive Teaching Approaches* and *Modeling Positive Behavior by Example* show lower or less consistent agreement. This pattern suggests that these topics are harder to distinguish consistently from neighboring topics.

Because the mapping from code clusters to final topics involves researcher judgment, persistent overlap across topics can be addressed by moving to a more aggregated topic structure, a granularity-reliability trade-off. Given the lack of reliability in two of the topics from the six-topic model, we iterated back to Step 5 and devised a four-topic model instead. We repeat the human and machine classification and inter-rater reliability analysis, with results reported in Table 6. Agreement improves substantially under the consolidated four-topic framework. Spearman's ρ exceeds 0.80 for all rater pairs and all Cohen's κ values are at least 0.60, placing them at or very near the conventional threshold for substantial

agreement. These improvements suggest that the four-topic structure reduces ambiguity and yields more reliable classifications. We therefore use the four-topic framework as the final *classroom management* topic model.

Table 6: Inter-Rater Reliability in Rank-Based Four-Topic Classification

Raters:	**GPT&H1**	**GPT&H2**	**H1&H2**
Per-Topic Cohen κ (Ranks)			
Authoritative Approach for Structure and Accountability	0.749	0.834	0.742
Student-Centered Engagement for Empowering Relationships	0.668	0.603	0.608
Positive Reinforcement Strategies and Systems	0.835	0.797	0.788
Modeling Positive Behavior by Example	0.605	0.621	0.679
Overall Rank-Order Agreement			
Spearman Rank Correlation ρ	0.817	0.836	0.817

Note: GPT = GPT-4o-mini. Human raters are denoted as H1 and H2, where H1 is the principal investigator and H2 is a Ph.D. student. Reliability is computed on the same random sample of 200 responses. Per-topic Cohen's κ is computed from raters' rank assignments for each topic. Overall Spearman rank correlation ρ summarizes agreement in the ordering of topic scores across the four-topic classification. Higher values of both Cohen's κ and Spearman ρ indicate greater agreement.

The four-topic model for *classroom management* is described in Table 7. Based on the underlying code clusters, codes, and texts rooted in prior steps, the research team finalizes the naming of the set of topics and their descriptions, which provide the substance of the document classification instructions for human or machine raters. The topic named *Authoritative Approach for Structure and Accountability* focuses on creating structured learning environments with explicit rules and expectations. Further description of this topic and the other three is in the table. This is the most prevalent topic in 45 percent of response documents, and 35 percent of the average response focuses on this topic.

Table 7: Four-Topic Model of *classroom management* Question Description and Prevalence

ID: Theme	Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
1: Authoritative Approach for Structure and Accountability	Structured learning environments with explicit rules and expectations, including firm but fair consequences. Teacher sets non-negotiable routines during the first classes, yet may invite limited but relevant student input when creating rules. Emphasizes accountability, disciplinary actions or plans, corrective measures at the group or student level, and consistently high standards.	45%	0.35
2: Student-Centered Engagement for Empowering Relationships	Teachers design lively, inclusive lessons that invite every learner to participate—through debates, hands-on projects, and other group or class-wide activities that elevate student voice. They flexibly tailor support to individual interests and emotional needs, building trust and a sense of belonging while encouraging autonomy and shared responsibility. They generate respectful relationships through open communication and empathy, creating nurturing learning environments where every learner feels included and valued.	45%	0.42
3: Positive Reinforcement Strategies and Systems	Reinforces desired behaviors with praise, points, token economies, and other tangible rewards; feedback is immediate and motivational. Emphasizes prevention rather than punishment and rarely mentions consequences.	7%	0.09
4: Modeling Positive Behavior by Example	Teachers intentionally model the attitudes and behaviors they expect from their students. Rather than relying on rules or rewards, they lead by example, demonstrating these qualities through their own actions. Teachers often emphasize their own personal conduct and attitudes as a tool to reinforce the desired norms.	2%	0.11

Note: Question text is "*Describe your classroom management style. Why have you found this style to be effective?*" Dmnt Tpc is Dominant Topic. Avg Scr is Average Score.

As an additional validation exercise, we inspect the highest-scoring responses within each topic to determine whether they align with our interpretation of clear cases. Appendix Table A.22 presents these examples for readers to audit. The top-rated responses align closely

with the intended meaning of each topic, providing additional support for the classification framework.

4 What Are Applicant Responses About?

4.1 Thematic Analysis

In the previous section we describe the HLTM procedure for the *classroom management* question. We apply this HLTM process to applicant responses for each of the 12 questions, though we do not repeat the full workflow for every question. Analogous to Table 7, Table A.7 presents the final topic models and summary statistics for each of the other questions. They show both what applicants tend to discuss in response to each question and how strongly different topics appear across responses. To illustrate what the procedure yields without offering a full substantive interpretation of all 12 questions, we highlight several question blocks below.

For the progress-tracking question (*student progress tracking*), the dominant topics revolve around concrete assessment routines. The most prevalent topics focus on instructional assessment and feedback practices (Topic 5) and data-driven assessment (1), where candidates describe quizzes, tests, exit tickets, and other checks for understanding as mechanisms for monitoring learning. Less common but clearly separable topics include individualized progress monitoring (3), holistic indicators such as engagement or self-regulation (4), and the use of specific tracking tools or documentation systems (6). Even within a single question, the method distinguishes between *what* teachers use to judge progress, such as tests or observational indicators, and *where* that information is recorded, such as assessment routines or record-keeping systems, and attaches prevalence estimates to each response.

The question asking how teachers build a classroom culture in which all students feel valued (*students feel valued*) has a different but equally structured profile. The most common topic centers on relational and engaging practices for inclusive classrooms (its Topic 3), in which candidates emphasize trust, open communication, and active participation as mechanisms for helping students feel known and valued. Additional topics capture culturally inclusive practices that affirm diversity and identity (5) and positive reinforcement and recognition of student growth (6). Other topics reflect high expectations and predictable structures (1), modeling respectful behavior (2), and personalized, student-centered instruction (4). Together, these topics show that when applicants are asked directly about inclusion, they describe a mix of relational work, cultural responsiveness, and instructional differentiation. The method assigns each strand a measurable prevalence in each response.

Questions about improvement and learning philosophy show how the approach generalizes to more abstract content. The improvement question (*difficult feedback*) yields topics around self-reflection (its Topic 1), adaptive teaching methods (2), and collaborative communication and relationship building (3). These topics capture, respectively, how candidates describe reflecting on their practice, changing instruction in response to feedback or student signals, and situating professional growth within collegial or relational contexts. The learning-styles and philosophy questions (*accommodate learning styles*, *philosophy learning growth*, and *learning philosophy*) show related but distinct emphases on individualized support (*accommodate learning styles*.4), differentiated and data-driven methods (*accommodate learning styles*.5), and personalized or student-centered differentiation (*philosophy learning growth*.1; *learning philosophy*.1 and *learning philosophy*.4). They also include separate topics focused on classroom climate and relationships (*philosophy learning growth*.2; *learning philosophy*.2). In these cases, the method separates the tailoring of content, pacing, and instructional supports from broader claims about community and belonging, again with associated prevalence measures for each applicant response.

The feedback-source question (*feedback sources*) illustrates how the same procedure applies to inputs and influences outside the classroom. The most common topic foregrounds formal mentorship and supervisory feedback (its Topic 1), consistent with evaluation and coaching structures in schools. The model also identifies distinct topics describing peer collaboration (3), multi-source stakeholder feedback (4), reflective uptake of feedback (6), and student responses as real-time instructional feedback (5). The resulting profile shows not only where candidates say they receive feedback, but also whether they frame feedback primarily as an institutional process, a collegial exchange, a multi-sided dialogue with families and students, or an individual habit of reflection.

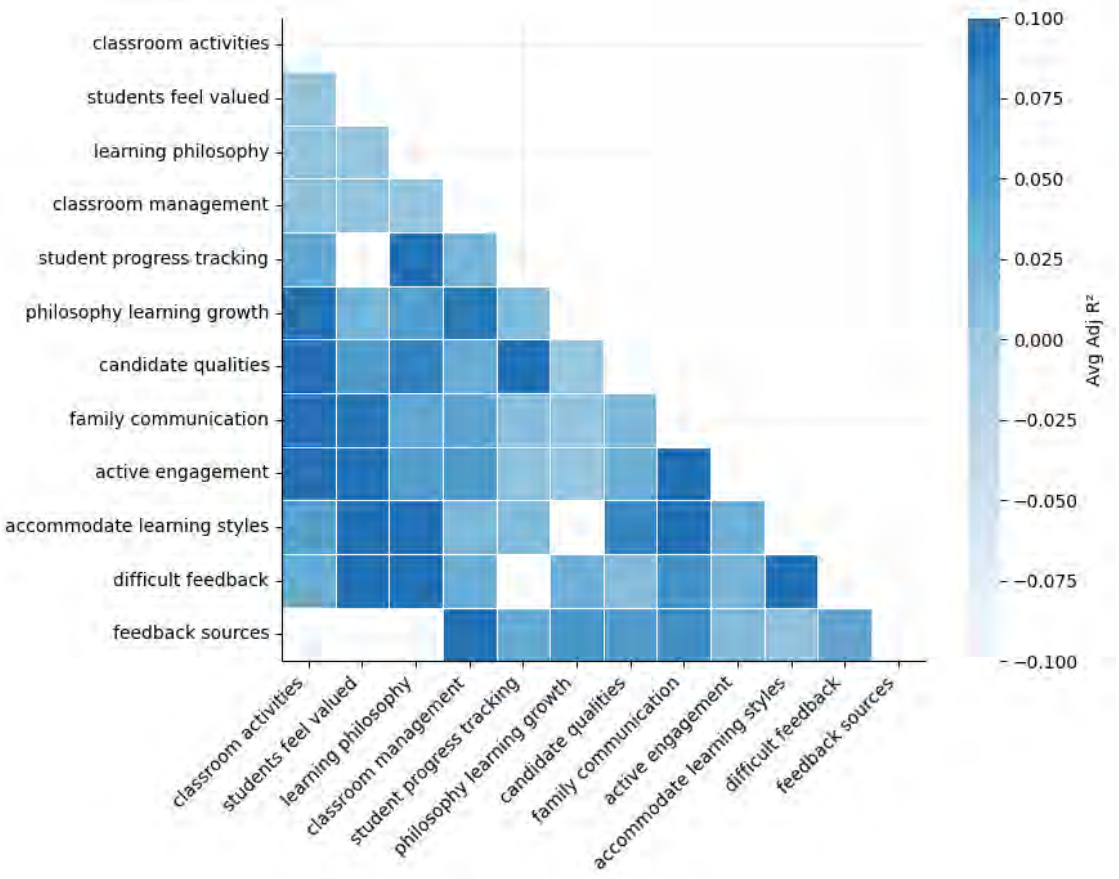
A full substantive interpretation of the topics for each of the 12 questions is beyond the scope of this paper as each question could support a separate qualitative analysis, but Table 7 provides an illustrative example, documenting the output of the text-to-data procedure and providing an interpretable summary of the information in a sample question. Researchers interested in the topics covered by these questions, like how early-career teachers talk about creating an inclusive environment, what they do to improve, how they think about state standards, or what they expect from administrators, can use Table A.7 to see the final topic models for all questions and ground theory or as an input for future qualitative work. For our purposes, the key point is that the same procedure yields interpretable, transparent, and auditable question-specific profiles that preserve substantive content while making written interview responses available for quantitative analysis.

4.2 Analysis of Information Overlap across Questions

When choosing which questions to ask, it is important to know whether different questions elicit similar or distinct dimensions of candidates' job-relevant characteristics. To assess overlap across questions, we examine how topic scores co-vary across pairs of questions among applicants who answered both questions in a pair. For each pair, we regress each topic score from question A on the vector of topic scores from question B. These regressions provide a descriptive measure of cross-question association in topical emphasis.

Because topic scores for a given question are normalized to sum to one, we omit one topic from the predictor vector to avoid perfect multicollinearity. We also omit one outcome topic from each question to avoid reporting mechanically redundant equations. For the *classroom management–students feel valued* pair, for example, we omit *classroom management*.3 and *students feel valued*.4 as outcomes, and we omit one topic from each question as a predictor. The resulting coefficients can be interpreted as partial associations between topics across questions, conditional on the remaining included topic scores from the predictor question. To summarize these many pairwise relationships, we present the heatmap in Figure 4 representing the average R^2 from the cross-question topic regressions for each pair of questions. The heatmap shows that topical overlap is minimal, suggesting that the questions elicit distinct dimensions of candidate responses.

Figure 4: Semantic overlap across pairs of questions



Note: Each cell reports the average adjusted R^2 across topic-level cross-question regressions for a pair of questions, estimated using applicants who answered both. For each regression, a topic score from one question is regressed on the vector of topic scores from the other question, with one predictor topic omitted. Higher values indicate greater predictive association in topical content across questions and greater information overlap.

4.3 A Validity Check

To provide a more interpretable validity check for the topic measures and classifications, we focus on one pair of questions: *classroom management* and *students feel valued*. Two authors independently reviewed the topic descriptions and identified pairs of topics they expected, on theoretical grounds, to be positively associated across the two questions. The expectation is that conceptually related topics should display larger and more consistently positive associations than conceptually unrelated topics. We restrict attention to positive associations because negative coefficients are difficult to interpret in this setting. Since topic weights within a question must sum to one, a positive association between one predictor and one outcome can mechanically induce negative coefficients for other topics.

The authors then compared their lists and agreed on a set of hypothesized positive topic

pairs. These pairs are denoted in **bold** in Table 8; all other topic pairs appear in plain text. For example, *classroom management*.1 (Authoritative Approach for Structure and Accountability) and *students feel valued*.1 (High Expectations and Predictable Structures for Student Success) were expected to be positively related because both emphasize accountability, routines, and predictable structures.

Table 8 reports the regression coefficients, with outcomes listed in rows and predictors listed in columns. In the regression where *classroom management*.1 is the outcome, the coefficient on *students feel valued*.1 is 0.20 and statistically significant at the 1 percent level, consistent with the hypothesized association. This coefficient implies that applicants who allocate 1 percentage point more of their *students feel valued* response[21] to *students feel valued*.1 (High Expectations and Predictable Structures for Student Success) tend, on average, to allocate 0.2 percentage points more of their *classroom management* response[22] to *classroom management*.1 (Authoritative Approach for Structure and Accountability), conditional on the other included *students feel valued* topic scores. Because both topics emphasize accountability to structures, this pattern suggests that the topic scores capture a stable aspect of candidates' responses across related questions. *Classroom management*.1 is also significantly associated with *students feel valued*.2, although we did not hypothesize that relationship. Coefficients on *students feel valued*.3, *students feel valued*.4, and *students feel valued*.5 are small and not statistically significant, consistent with the absence of a hypothesized association.

In total, Table 8 reports 30 estimated coefficients. Eight are redundant by symmetry, leaving 22 distinct pairwise relationships between topics across the two questions. Five of these relationships were hypothesized to be positive; they appear as 8 coefficients in the table because some relationships are observed in both regression directions.[23] Four of the 5 hypothesized relationships are positive and statistically significant at the 10 percent level. Under a null of no association, one would expect fewer than one statistically significant coefficient at this threshold among eight tests, although these tests are not independent. Among the remaining 17 non-redundant relationships that were not hypothesized to be positive, only one is positive and statistically significant.[24] These patterns provide supportive evidence that the topic scores behave in expected ways across related questions and capture meaningful variation in applicant responses.

[21] *How do you build a classroom culture in which all students feel valued?*

[22] *Describe your classroom management style. Why have you found this style to be effective?*

[23] The symmetric coefficients differ because the regressions condition on different sets of predictors.

[24] Several other coefficients are negative and statistically significant. We do not interpret them as evidence against validity. With linearly dependent topic shares, a strong positive relationship for one predictor can mechanically produce negative coefficients for others.

Table 8: Regression-Based Associations between Topic Scores across Question Pairs Answered by the Same Applicant

$Y_i \setminus X_j$	classroom management.1	classroom management.2	classroom management.3	students feel valued.1	students feel valued.2	students feel valued.3	students feel valued.4	students feel valued.5
classroom management.1				**0.20*** **	0.10***	0.03	0.00	0.02
classroom management.2				-0.10***	-0.08***	**0.05***	0.04	**0.04***
classroom management.4				-0.06***	**0.01**	-0.01	0.02	-0.03*
students feel valued.1	**0.03*** **	-0.03***	-0.01					
students feel valued.2	0.00	-0.05***	-0.03***					
students feel valued.3	0.00	**0.03***	-0.01					
students feel valued.5	0.00	**0.04***	0.00					
students feel valued.6	0.01	0.01	**0.07*** **					

Note: Entries are coefficients from regressions of the row topic score Y_i on the column topic score X_j and the other included topic scores from the same predictor question. Bold entries indicate topic pairs hypothesized to be positively associated before estimating the regressions. One topic from each question is omitted from the predictor vector to avoid perfect multicollinearity, and one outcome topic from each question is omitted to avoid reporting mechanically redundant equations. * Significant at 10%, *** significant at 1%.

5 What Are the Measurement Properties of Applicant Responses to Open-Ended Questions?

A central objective of eliciting applicant responses to open-ended questions is to generate information that may help predict future job performance and/or retention. Responses with the measurement properties we describe below are more likely to be useful for that purpose, a point we return to in the conclusion. Good interview questions should generate job-relevant variation in responses and differentiate candidates along dimensions plausibly related to performance in the role. They should also be less susceptible to strategic or scripted answers aimed at matching perceived interviewer expectations, because questions with an "obvious best" answer compress the distribution of responses and reduce responses' informational

content. Questions that elicit highly similar answers provide limited information for differentiating applicants and may be more vulnerable to checklist-like responses.

Because interview instruments inform employment decisions, they should also be designed and interpreted in ways that minimize the risk of systematic disparities across legally protected groups. The subsections that follow examine how topic scores vary across candidate characteristics and provide question-level diagnostics that summarize each item's candidate-differentiating signal and flag cases where topic emphasis, if used incautiously, could be associated with disparate selection across protected characteristics.

5.1 Properties of Responses and Applicant Correlates

This section provides question-level summary statistics and diagnostics that describe how applicant responses vary across topics and respondent characteristics, with an emphasis on their potential practical use.

We examine how topic classifications vary across applicant characteristics. For each question, the outcome is the composition of topic weights assigned to each response. Because the weights assigned to topics within a response sum to one, we model their log-ratios using an additive log-ratio (ALR) specification that treats one topic as the compositional baseline.[25] Results are presented in Appendix A.10 and reveal that topic emphasis is associated with applicant characteristics for nearly every question. For example, in the classroom-management question, female applicants place greater relative weight on the Authoritative topic (1) and the Positive-Reinforcement topic (3) relative to the baseline topic (4).[26]

[25]Formally, for the *classroom management* question, we define

$$y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})',$$

where y_{ij} denotes the score on classroom management topic j (topics 1–4), and estimate:

$$\log \frac{y_{ij}}{y_{i4}} = \mathbf{x}_i' \beta_j + \varepsilon_{ij}, \qquad (j = 1, 2, 3),$$

Topic *classroom management. 4* serves as the denominator, and $\mathbf{x}_i$ includes indicators for sex, race/ethnicity, and experience categories. Each equation is estimated via OLS with cluster-robust standard errors at the applicant level. Because log ratios are undefined when a topic weight equals zero, zero values in the assigned topic probabilities were replaced by small constants $\delta \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ and re-normalized; results were virtually identical across values. For completeness, we also estimated a Dirichlet regression using the mean–precision parameterization, which yielded similar qualitative patterns.

[26]To our knowledge, the closest comparison in the existing literature is Barni et al. (2018), who study teachers' self-reported classroom-management styles using a compressed framework with three categories—authoritative, authoritarian, and permissive. They find that gender is only weakly related to classroom-management style overall, with one notable exception: female teachers are more likely to report an authoritative style. We interpret these associations cautiously. Because measures depend on applicants' written self-descriptions in a high-stakes hiring setting, they may capture both underlying classroom-management approach and also differences in self-presentation, which has been shown to differ systematically by gen-

Race and ethnicity are also associated with topic emphasis. Applicants who self-identify as *Black* place greater relative weight on Authoritative (1) and Student-Centered practices (2), whereas applicants who self-identify as *Asian* are relatively less likely to describe their classroom management style in terms aligned with the Authoritative topic (1). *Experience* is also associated with responses: fully novice teachers (0 years) are less likely to emphasize authoritative management (1) than the veteran reference group ($>$ 10 years), while intermediate experience groups (1 to 10 years) do not differ appreciably from the veteran reference group after controlling for sex and race/ethnicity. This suggests that novice teachers are less likely to start their careers emphasizing structure in their written responses, but tend to change their opinions quickly after. Because these estimates are cross-sectional and based on high-stakes self-descriptions, we interpret them as associations in how applicants present their classroom-management approach, not as evidence of within-teacher changes in pedagogical beliefs.

Because gender, race and ethnicity are correlated with a screening instrument, it is necessary to examine risks for adverse selection. To do this, we compute the share of extreme selection ratios, which summarizes an adverse-impact/disparate-treatment screening exercise by question and topic. For each of a question's K topics, we compute implied demographic selection ratios under stylized rules in which selection is based solely on applicants' highest or lowest scores on that topic. We consider four demographic contrasts: Black/White, Hispanic/White, Asian/White, and Female/Male. This produces $8K$ tail–contrast–topic selection ratios per question. We flag ratios below 0.8 or above 1.25 as extreme. The reported proportion is the share of these $8K$ ratios flagged as extreme. A higher share means that more topics within a question could raise concerns about disparate selection if used in isolation. This is not evidence of legal violation on its own: real hiring decisions need not turn on a single topic from a single question. Rather, the measure identifies questions and topics where caution is warranted when interpreting or operationalizing topical focus.

Table 9 reports two diagnostics. The first reports how often a stylized topic-based selection rule would produce extreme demographic selection ratios. The second, across-response dispersion, summarizes how much topic mixtures vary across applicants. Together, these diagnostics speak to whether topic emphasis could raise disparate-selection concerns if used incautiously in hiring, and whether a question helps differentiate candidates. The first concern is substantively important because, as shown above, topic weights are significantly

der (Exley and Kessler, 2022). While backlash-avoidance arguments often predict incentives for women to present themselves as less assertive (Rudman and Glick, 2001; Moss-Racusin and Rudman, 2010), that prediction may not extend cleanly to female-majority occupations such as teaching, where male applicants may in some cases be evaluated less favorably (Fullard, 2025).

associated with applicant demographic characteristics.[27]

Table 9: Prevalence of Extreme Implied Selection Ratios and Dispersion of Topic Mixtures by Question

Question	**Share of Extreme Selection Ratios**	**Across-Response Dispersion**	N	K
classroom activities	0.29	365.76	6296	6
students feel valued	0.21	272.11	5896	6
learning philosophy	0.25	169.88	4770	4
classroom management	0.22	180.76	4629	4
student progress tracking	0.23	332.96	1910	6
philosophy learning growth	0.46	74.60	1622	3
candidate qualities	0.40	239.52	1465	5
family communication	0.31	315.69	1342	6
active engagement	0.33	288.85	1164	5
accommodate learning styles	0.53	227.96	817	5
difficult feedback	0.41	101.97	664	4
feedback sources	0.33	281.65	434	6

Note: The Share of Extreme Selection Ratios is the proportion of implied demographic selection ratios (Black/White, Hispanic/White, Asian/White, and Female/Male) that are extreme, falling either below 0.8 or above 1.25. These rations are computed under stylized rules by comparing hiring outcomes for applicants with the highest or lowest scores on each topic. For each question, there are $8 \times K$ implied ratios: 4 demographic contrasts times 2 directions of selection (highest or lowest topic scores) times K topics. These diagnostics are descriptive flags and do not represent actual hiring outcomes. Across-Response Dispersion is the isometric log-ratio (ilr) total variance, defined as the trace of the covariance matrix of ilr coordinates, and reflects heterogeneity in topic mixtures across responses. Zeros in compositions are replaced with a small ε, and vectors are re-closed before log-ratio transforms.

Each row pools all responses to a given question and treats each response as a composition over K topics, with topic shares p_k summing to one. The diagnostics summarize, for each question, (i) how frequently a stylized "hire-on-this-topic" rule would produce large demographic imbalances in selection, and (ii) how much applicants differ from one another in topical emphasis.

The selection-ratio diagnostic highlights a different dimension of question performance. *Accommodate learning styles* has the largest share of extreme implied selection ratios (0.53), followed by *philosophy learning growth* (0.46), *difficult feedback* (0.41), and *candidate qualities* (0.40). These questions therefore warrant greater caution if topic emphasis were used directly in screening. At the other end, *students feel valued*, *classroom management*, and *student*

[27]Appendix A.9 reports additional dispersion diagnostics, including within-response Gini–Simpson evenness, Aitchison concentration, the prevalence of near-zero topic weights, and the prevalence of monothematic responses.

progress tracking have lower shares of extreme implied selection ratios (0.21–0.23).

The dispersion column, across-response isometric log-ratio (ilr) total variance, captures how much topic mixtures vary across responses to a given question. Higher values indicate that applicants emphasize different topics to different degrees, suggesting a greater candidate-differentiating signal. Lower values indicate that applicants' topical response profiles are more similar to one another. For interview design, across-response dispersion is especially important: a question that yields little variation across applicants provides a limited basis for differentiating candidates, regardless of whether uniformity reflects benign consensus or a more concerning "obvious right answer" dynamic. Conversely, higher across-response dispersion suggests the question elicits differences in how candidates frame their experiences or priorities.

Turning to the results, the dispersion measures point to two broad response profiles and highlight which questions appear weaker or stronger as candidate differentiators. *Philosophy learning growth* and *difficult feedback* have the lowest across-response heterogeneity (ilr $\approx$ 75–102). In practical terms, candidates tend to cover topics similarly to one another. This pattern is consistent with responses that may be partially scripted or checklist-like and provide a more limited basis for differentiating candidates, although it could reflect either benign consensus or the nature of broad reflective prompts.

A second set of questions is distinguished by much larger across-response dispersion. *Classroom activities*, *student progress tracking*, and *family communication* show the highest ilr variances ($\approx$ 316–366), indicating substantial cross-candidate heterogeneity in topical emphasis and therefore a stronger candidate-differentiating signal. *Active engagement* also shows relatively high across-response dispersion (ilr $\approx$ 289). By contrast, *learning philosophy*, *classroom management*, *candidate qualities*, and *accommodate learning styles* fall in the middle range, suggesting some differentiating signal but less than the highest-dispersion questions. Overall, Table 9 separates questions that more clearly surface differences in candidate emphasis from questions that provide weaker differentiation or raise more frequent disparate-selection flags.

6 Conclusion

This study is one of the first to operationalize the analysis of written interview-style short-answer questions in teacher screening. We develop a transparent method for topic discovery that extracts question-specific topics from applicant responses in a scalable manner while preserving interpretability and auditability. We leverage a scalable approach for scoring any set of responses against discovered topics. These measures provide a transparent input for

validation, monitoring, and local decision-making.

We provide new, grounded theory for understanding each question and responses to it. Practitioners interested in studying any of these questions or the discovered topics can adapt the prompts in Appendix A.6 and populate them with the topic definitions in Table A.7. This will generate response-level topic measures that can be used to audit the content of applicant responses, compare questions, and evaluate whether particular questions elicit useful variation.

The analysis of inter-question topic associations suggests that applicants' responses are related across questions in conceptually expected ways, providing supportive evidence for the validity of the HLTM-ITA framework and classification approach. We provide evidence that LLM-based scoring of responses along the discovered topics performs similarly to human scoring. We also provide quantitative diagnostics of how topic mixtures vary within and across applicant responses using real high-stakes teacher application data. These diagnostics speak to the candidate-differentiating value of different questions as a selection tool and to the concern that answers to high-stakes interview questions may be strategically tailored to perceived expectations. Most questions generate meaningful variation in topic emphasis across applicants, which is inconsistent with complete convergence on a single widely known script. However, some questions, especially broad reflection or philosophy prompts, produce more uniform and balanced response profiles, suggesting that they may be less useful for differentiating candidates. These patterns do not rule out strategic responding, but provide a practical way to identify questions that appear informative.

Finally, because this work speaks directly to the use of algorithmic tools in hiring, we provide an expanded appendix for practitioners that reports associations between applicant characteristics and response topics, as well as summary measures indicating where topic-based selection rules could produce extreme demographic selection ratios. Under Title VII of the Civil Rights Act, employers may not engage in intentional discrimination (disparate treatment) and must also avoid neutral practices that *cause* disparate impact unless the practice is *job-related and consistent with business necessity*; recent EEOC technical assistance makes clear that this analysis applies to AI and algorithmic selection tools used to make or inform employment decisions, and that employers remain responsible for vendor tools they deploy (U.S. Equal Employment Opportunity Commission, 2023, 2024). The federal *Uniform Guidelines on Employee Selection Procedures* (UGESP) provide the widely used "four-fifths rule"—a screening heuristic, not a safe harbor or a cap on allowable bias—under which a selection rate for any protected group below 80 percent of the highest group's rate is generally regarded as evidence of adverse impact (U.S. Equal Employment Opportunity

Commission et al., 1978; Legal Information Institute, Cornell Law School, 1978).[28]

To this end, our diagnostics report the share of topics within each question that would trigger the four-fifths flag under stylized topic-based selection rules, leading to inadvertent disparate treatment if used without due caution. These diagnostics are not meant to serve as stand-alone hiring rules not are they legal conclusions, but they help identify where additional validation, monitoring, and caution would be warranted before using response topics as part of a broader selection system.

By converting short-answer responses into quantitative topic measures, the framework enables the efficient evaluation of the information elicited by these questions and, in future work, the ability to assess their predictive validity for hiring, retention, and performance outcomes. All this helps illuminate which questions districts should prioritize and which kinds of responses they should screen for.

References

Backes, Ben, Lauren Covelli, Michael DeArmond, Elise Dizon-Ross, Dan Goldhaber, Julia Kaufman, and Umut Özek (2024) "More than Just Adding Courses: Evidence on Algebra and Equity from the American Mathematics Educator Study," *Calder Research Brief No. 39, National Center for Analysis of Longitudinal Data in Education Research.*

Barni, Daniela, Chiara Russo, and Francesca Danioni (2018) "Teachers' Values as Predictors of Classroom Management Styles: A Relative Weight Analysis," *Frontiers in Psychology*, 9, 1970, 10.3389/fpsyg.2018.01970.

Bartanen, Brendan, Andrew Kwok, Andrew Avitabile, and Brian Heseung Kim (2025) "Why do you want to be a teacher? A natural language processing approach," *Educational Researcher*, 54 (1), 7–20.

Boyatzis, Richard E (1998) *Transforming Qualitative Information: Thematic Analysis and Code Development*: Sage.

[28]Several jurisdictions also impose additional obligations on automated decision tools. New York City's Local Law 144 requires an annual bias audit, a public summary, and candidate notice before using an automated employment decision tool (New York City Department of Consumer and Worker Protection, 2023a,b). Illinois regulates AI-assisted video interviews, including notice/consent and data-handling duties (Illinois General Assembly, 2019). Colorado's AI Act (SB 24-205) imposes risk-management and impact-assessment duties for high-risk AI systems, including employment decisions, effective in 2026 (Colorado General Assembly, 2024). In addition, ADA enforcement guidance cautions that algorithmic tools must not screen out qualified individuals with disabilities and may require reasonable accommodation in the assessment process (U.S. Equal Employment Opportunity Commission, 2022; U.S. Department of Justice, Civil Rights Division, 2022).

Braun, Virginia and Victoria Clarke (2006) "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology*, 3 (2), 77–101.

Byrne, David and Dr Aiden Carthy (2021) "A Qualitative Exploration of Post-primary Educators' Attitudes Regarding the Promotion of Student Wellbeing," *International Journal of Qualitative Studies on Health and Well-Being*, 16 (1), 1946928.

Cascio, M Ariel, Eunlye Lee, Nicole Vaudrin, and Darcy A Freedman (2019) "A Team-Based Approach to Open Coding: Considerations for Creating Intercoder Consensus," *Field methods*, 31 (2), 116–130.

Chall, Jeanne Sternlicht and Edgar Dale (1995) *Readability Revisited: The New Dale–Chall Readability Formula*: Brookline Books.

Chen, et al. (2024) "A Computational Method for Measuring "Open Codes" in Qualitative Analysis," *ACM Transactions on Human-Computer Interaction*, 12, 87–102.

Chetty, Raj, John N Friedman, and Jonah E Rockoff (2014) "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104 (9), 2633–2679.

Colorado General Assembly (2024) "SB 24-205: Consumer Protections for Artificial Intelligence," `https://leg.colorado.gov/bills/sb24-205`, Effective Feb. 1, 2026, for high-risk AI systems.

De Paoli, Stefano (2024) "Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach," *Social Science Computer Review*, 42 (4), 997–1019, 10.1177/08944393231220483.

Donaldson, Bryan Phillip (2024) *Consumer Product Goods Employability Qualifications for Recent Graduates in Business: A Qualitative Descriptive Approach* Ph.D. dissertation, Grand Canyon University.

Dunlop, Patrick D., Djurre Holtrop, and Serena Wee (2022) "How Asynchronous Video Interviews Are Used in Practice: A Study of an Australian-based AVI Vendor," *International Journal of Selection and Assessment*, 30 (3), 448–455, 10.1111/ijsa.12372.

Exley, Christine L. and Judd B. Kessler (2022) "The Gender Gap in Self-Promotion," *The Quarterly Journal of Economics*, 137 (3), 1345–1381, 10.1093/qje/qjac003.

Flesch, Rudolf (1948) "A New Readability Yardstick," *Journal of Applied Psychology*, 32 (3), 221–233, 10.1037/h0057532.

Fullard, Joshua (2025) "Is the School Workforce Biased Against Men?" *Economics Letters*, 251, 112337, 10.1016/j.econlet.2025.112337.

Grootendorst, Maarten (2022) "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," *arXiv preprint arXiv:2203.05794*.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020) "spaCy: Industrial-strength Natural Language Processing in Python," 10.5281/zenodo.1212303.

Hopkins, Daniel J. and Gary King (2010) "A Method of Automated Nonparametric Content Analysis for Social Science," *American Journal of Political Science*, 54 (1), 229–247, 10.1111/j.1540-5907.2009.00428.x.

Illinois General Assembly (2019) "Artificial Intelligence Video Interview Act, Public Act 101-0260," `https://www.ilga.gov/Legislation/publicacts/view/101-0260`.

Jackson, C Kirabo, Jonah E Rockoff, and Douglas O Staiger (2014) "Teacher effects and teacher-related policies," *Annu. Rev. Econ.*, 6 (1), 801–825.

Jackson, Jane (2018) *Interculturality in International Education*: Routledge.

Jalali, Mohammad S. and Ali Akhavan (2024) "Integrating AI Language Models in Qualitative Research: Replicating Interview Data Analysis with ChatGPT," *System Dynamics Review*, 40, e1772, 10.1002/sdr.1772.

Kincaid, J. Peter, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom (1975) "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," Research Branch Report 8–75, Naval Technical Training Command, Research Branch.

Landis, J. Richard and Gary G. Koch (1977) "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 33 (1), 159–174, 10.2307/2529310.

Langer, Markus et al. (2024) "A Quasi-Experimental Investigation of Differences between Face-to-Face and Videoconference Interviews in an Actual Selection Process," *Applied Psychology*, 10.1111/apps.12558.

Legal Information Institute, Cornell Law School (1978) "29 C.F.R. §1607.4 — Information on Impact," Online reproduction of the CFR text and notes, https://www.law.cornell.edu/cfr/text/29/1607.4.

Levashina, Julia, Christopher J. Hartwell, Frederick P. Morgeson, and Michael A. Campion (2014) "The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature," *Personnel Psychology*, 67, 241–293, 10.1111/peps.12052.

Liu, Yujia, Emily K Penner, Sabrina Solanki, and Xuehan Zhou (2025) "What Educator Applicants' Short Essays Tell You about Their Potential Job Fit and Performance," *Journal of Education Human Resources*, 43 (2), 275–364.

Long, et al. (2024) "Evaluating Large Language Models in Analyzing Classroom Dialogue," *npj Science of Learning*, 9, 60, 10.1038/s41539-024-00273-3.

Macan, Therese (2009) "The Employment Interview: A Review of Current Studies and Directions for Future Research," *Human Resource Management Review*, 19 (3), 203–218.

McKinney, Wes (2010) "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, 51–56.

Melão, Nuno and João Reis (2021) "Social Networks in Personnel Selection: Profile Features Analyzed and Issues Faced by Hiring Professionals," *Procedia Computer Science*, 181, 42–50.

Moss-Racusin, Corinne A. and Laurie A. Rudman (2010) "Disruptions in Women's Self-Promotion: The Backlash Avoidance Model," *Psychology of Women Quarterly*, 34 (2), 186–202, 10.1111/j.1471-6402.2010.01561.x.

New York City Department of Consumer and Worker Protection (2023a) "Automated Employment Decision Tools (AEDT)," https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page, Program page, enforcement and resources.

——— (2023b) "Automated Employment Decision Tools: Frequently Asked Questions," June, https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn (2015) "The Development and Psychometric Properties of LIWC2015,"Technical report, University of Texas at Austin, 10.15781/T29G6Z.

Posthuma, Richard A, Frederick P Morgeson, and Michael A Campion (2002) "Beyond Employment Interview Validity: A Comprehensive Narrative Review of Recent Research and Trends over Time," *Personnel Psychology*, 55 (1), 1–81.

Potočnik, Kristina, Neil Anderson, Marise Born, Martin Kleinmann, and Ioannis Nikolaou (2021) "Paving the Way for Research in Recruitment and Selection: Recent Developments, Challenges and Future Opportunities," *European Journal of Work and Organizational Psychology*, 30 (2), 159–174, 10.1080/1359432X.2021.1904898.

Rockoff, Jonah E, Brian A Jacob, Thomas J Kane, and Douglas O Staiger (2011) "Can you recognize an effective teacher when you recruit one?" *Education finance and Policy*, 6 (1), 43–74.

Rubin, Jacob and Jason Grissom (2026) "Leveraging Large Language Models to Assess Short Text Responses," January.

Rudman, Laurie A. and Peter Glick (2001) "Prescriptive Gender Stereotypes and Backlash Toward Agentic Women," *Journal of Social Issues*, 57 (4), 743–762, 10.1111/0022-4537.00239.

Sajjadiani, S., A. Sojourner, J. Kammeyer-Mueller, and E. Mykerezi (2019) "Using machine learning to translate applicant work history into predictors of performance and turnover," *Journal of Applied Psychology*, 104 (10), 1207–1225.

U.S. Department of Justice, Civil Rights Division (2022) "Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring," May, `https://www.ada.gov/resources/ai-guidance/`.

U.S. Equal Employment Opportunity Commission (2022) "The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees," technical assistance document, `https://www.eeoc.gov/eeoc-disability-related-resources/artificial-intelligence-and-ada`, Landing page that links to the 2022 ADA/AI technical assistance; EEOC has changed deep URLs.

——— (2023) "Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," technical assistance document, `https://data.aclum.org/storage/2025/01/EOCC_www_eeoc_gov_laws_guidance_select-issues-assessing-adverse-impact-software-algorithms-and-artificial.`

pdf, Official page has been reorganized; see EEOC Publications index. Stable copy of the document available via mirror link below.

——— (2024) "What is the EEOC's Role in AI?," April, https://www.eeoc.gov/sites/default/files/2024-04/20240429_What%20is%20the%20EEOCs%20role%20in%20AI.pdf.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978) "Uniform Guidelines on Employee Selection Procedures (1978)," Code of Federal Regulations, 29 C.F.R. Part 1607, https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607, Adopted jointly by EEOC, CSC, DOL, and DOJ.

Wang, Qile, Moath Erqsous, Kenneth E. Barner, and Matthew Louis Mauriello (2025) "LATA: A Pilot Study on LLM-Assisted Thematic Analysis of Online Social Network Data Generation Experiences," *Proceedings of the ACM on Human-Computer Interaction*, 9 (2), 1–28, 10.1145/3711022.

Woods, Megan, Trena Paulus, David P. Atkins, and Rob Macklin (2016) "Advancing Qualitative Research Using Qualitative Data Analysis Software (QDAS)? Reviewing Potential Versus Practice in Published Studies Using ATLAS.ti and NVivo, 1994–2013," *Social Science Computer Review*, 34 (5), 597–617.

Yeung, Matthew WL and Alice HY Yau (2022) "A Thematic Analysis of Higher Education Students' Perceptions of Online Learning in Hong Kong under COVID-19: Challenges, Strategies and Support," *Education and Information Technologies*, 27 (1), 181–208.

Zhang, He, Chuhao Wu, Jingyi Xie, ChanMin Kim, and John M Carroll (2023) "QualiGPT: GPT as an Easy-to-Use Tool for Qualitative Coding," *arXiv preprint arXiv:2310.07061*.

ONLINE APPENDIX

Short Answer

Aguilar, Goldhaber, Grout, Mykerezi, Pathak, & Sojourner

A Details on Hybrid LLM–Topic-Model Inductive Thematic Analysis (HLTM)

This appendix describes HLTM (HLTM-ITA) as implemented for the question: "Describe your classroom management style." It consists of four main steps: analytic pre-processing, response validation, topic labeling, and theme identification and refinement. Each step is detailed in the following sections.

A.1 Analytic Pre-processing

Data are at the question-response level consisting of an applicant-a's textual response to a particular question-q as part of their application for position-p with employer-e made at time-t: D_{eptqa}.

Pre-processing document text involves three steps—filtering out very short documents, duplicates or similar, and blank responses. These steps aim to improve the quality of topic discovery by reducing noise and focusing on meaningful variation in terms.

We discuss this process substantively with respect to one screening question: "Describe your classroom management style." We index this as Question *classroom management*. Some pre-processing details vary across screening questions but all follow a similar structure. Tables A.1 and A.2 provide details on choices made in the pre-processing stage and descriptive statistics for each question.

For the question *classroom management*, the raw dataset consisted of 5,068 applicant responses. The mean response contained 76 words, the median 57, and the standard deviation was 75. Responses ranged from empty (0 words) to a maximum of 938 words.

First, we filter out responses that are too short to contribute meaningful variation to the topic identification. To define a threshold for inclusion, we examined the shortest 5 percent of responses (less than 10 words) and the shortest 1 percent (less than 3 words). Ultimately, we excluded responses with fewer than 6 words, leaving 4,900 valid responses. These were most likely incomplete responses or data errors.

Secondly, we remove duplicate or nearly identical and blank responses, as these could skew topic distributions by introducing unnecessary noise or redundancy. Blank responses

were filtered out; any response without content was excluded. This step removed 0 additional records for this question. Similarly, repeated responses were examined, as identical or nearly identical answers across questions or interviews could bias topic distributions. Exact duplicates were identified (McKinney, 2010), removing 3 duplicate responses. Near-duplicates were identified using a similarity function between documents based on word embeddings (Honnibal et al., 2020). Responses flagged as near-duplicates with over 99 percent similarity were retained as single instances, reducing redundancy while preserving unique information, removing 118 more records from the *classroom management* responses. Table A.1 and A.2 report these statistics and resulting dropped documents from the pre-processing for Question *classroom management*, as well as the corresponding information for the other analyzed questions.

Table A.1: Descriptive Statistics of Raw Sample across Questions

Question	**Raw Count**	**Mean**	**Median**	**SD**	**Min**	**Max**
classroom activities	6789	67.36	47	72.71	0	1632
students feel valued	6387	78.75	59	76.50	0	1136
learning philosophy	5220	71.21	42	119.34	0	3732
classroom management	5068	76.49	57	75.07	0	938
student progress tracking	2028	59.76	39	60.59	0	522
philosophy learning growth	1847	113.99	76	160.76	0	4474
candidate qualities	1597	101.46	66	117.62	0	1205
family communication	1409	98.48	81	75.20	0	578
active engagement	1241	49.99	35	57.96	0	1062
accommodate learning styles	901	64.21	43	68.92	0	574
difficult feedback	931	117.87	87	111.67	0	979cv
feedback sources	477	55.47	37	72.77	0	1059

Note: All values calculated on raw sample. Mean, Median, SD, Min, and Max are the average, median, standard deviation, minimum, and maximum number of words, respectively, in the answers to each question.

Table A.2: Pre-processing and Validation of Responses to each Question: From Raw Counts to Final Validated Sample

Question	**Raw Count**	**Dropped From Raw Sample**				**Final Sample**
		Short	**Duplicates**	**Similar**	**Invalid**	
classroom activities	6789	4	27	98	364	6296
students feel valued	6387	48	15	234	209	5896

Continued on next page

Table A.2: Pre-processing and Validation of Responses to each Question: From Raw Counts to Final Validated Sample (continued)

Question	Raw Count	Dropped From Raw Sample				Final Sample
		Short	Duplicates	Similar	Invalid	
learning philosophy	5220	72	26	214	138	4770
classroom management	5068	168	3	118	150	4629
student progress tracking	2028	26	3	9	80	1910
philosophy learning growth	1847	29	1	155	40	1622
candidate qualities	1597	31	0	33	68	1465
family communication	1409	18	1	14	34	1342
active engagement	1241	27	1	6	43	1164
accommodate learning styles	901	21	0	5	58	817
difficult feedback	931	29	0	15	223	664
feedback sources	477	17	0	0	26	434

Note: All values correspond to the number of answers per question.

A.2 Response Validation

Before conducting the thematic analysis, we perform a final cleaning of the dataset to assess whether each response appropriately addresses the question. To achieve this, we use the OpenAI API to prompt a GPT model with three different sets of instructions: long, medium, and short. While all three sets serve the same purpose, they differ in the level of detail provided in the instructions.

The short instructions include the question text, the main task, the response text, and the answer ID. The medium-length instructions include the same elements as the short version but also incorporate a concise description of the classification criteria, specifying what qualifies as responsive or non-responsive. The long instructions provide the most detail, including the question text, the main task, the response text, the answer ID, a thoroughly detailed classification criterion defining responsiveness, and a set of example responses correctly classified as responsive or non-responsive.

Once all three classifications are completed, we merge the resulting datasets into a single file using the answer ID as the key. We then define three classification variables—long, medium, and short—corresponding to the different instruction sets. Given that each classification is based on a different set of instructions, variations may arise in what is considered responsive or non-responsive. To ensure inclusivity, we retain a response in the final dataset if it is classified as responsive in at least one of the three classifications. If a response is deemed non-responsive across all three instruction sets, we exclude it from the dataset. This process results in the removal of 150 responses, leaving a final dataset of 4,629 responsive

answers (Table A.2).

We use OpenAI's GPT-4o-mini to classify each response as either responsive (1) or non-responsive (0). This process is conducted on a per-instance basis, meaning that each response is evaluated through a separate API call.

The prompt for the "short" instructions is as follows:

```
def make_prompt(response, answer_id):
    prompt = f'''
    I need help looking for bad data. I have teacher responses to the question
    "Describe your classroom management style".
    Most responses seem like legitimate answers to the question prompt but some
    are data errors.
    Can you help determine which ones do not appear to be responsive to the
    question prompt as a reasonable person might answer this question.

    Response to classify:
    Answer ID: {answer_id}
    "{response}"

    Provide only the classification (1 or 0) with no extra text.
    '''
    return prompt
```

The prompt for the "medium"-length instructions is as follows:

```
def make_prompt(response, answer_id):
    """
    Creates a mid-sized prompt for GPT to classify a single response.
    """
    prompt = f"""
    Please classify the following teacher response to the question:
    "Describe your classroom management style."

    Classification criteria:
    - Classify as 1 (Responsive) if the response provides any meaningful insight
    into classroom management, including strategies, philosophies, or practices
    related to rules, engagement, structure, discipline, communication, or
    student relationships.
    - Classify as 0 (Non-Responsive) if the response is off-topic, generic,
    vague, or does not provide clear information about classroom management
```

```
    practices.

    Response to classify:
    Answer ID: {answer_id}
    "{response}"

    Provide only the classification (1 or 0) with no extra text.
    """
    return prompt
```

The prompt for the "long" instructions is as follows:

```
def make_prompt(response, answer_id):
    """
    Creates a prompt for GPT to classify a single response.
    """
    prompt = f"""
    Here is a teacher's response to the question:
    "Describe your classroom management style."

    Classify the response as 1 (Responsive) or 0 (Non-Responsive) based on the
    following criteria:

    1 (Responsive) if:
    - The response describes any aspect of a classroom management style,
    philosophy, or specific strategies.
    - The response includes engagement, rules, consequences, structure,
    discipline, relationships, fairness, communication styles, motivation,
    personal availability, student respect, or expectations.
    - If the response discusses how the teacher interacts with students to manage
    behavior, set expectations, or enforce structure, it should be classified as
    1.
    - If a teacher describes how they build relationships, communicate with
    students, set rules, maintain fairness, establish a structured environment,
    or prevent behavior issues, this still counts as 1.
    - Even if the response does not use technical terms like "discipline" or
    "classroom management," it should still be classified as 1 if it describes
    real classroom practices.
    - Even if the response is long, slightly unstructured, or conversational, it
    should be classified as 1 if it provides meaningful insight into classroom
```

```
management.

0 (Non-Responsive) if:
- The response is completely off-topic.
- The response says "See resume," "I prefer not to say," or provides a
generic phrase with no details.
- The response is vague or non-informative (e.g., "I have a style that works
for me").
- The response does not indicate any specific management approach,
philosophy, or structured strategy.

Examples:
1: "I use a structured approach with clear rules and emphasize student
engagement and fairness."
1: "My approach is direct, assertive, but also empathetic. I build
relationships with students and create engaging lessons to maintain
discipline."
1: "I set clear rules and expectations at the start of the school year and
make sure students understand the consequences."
1: "I communicate with students according to their personality and encourage
respect. I do not tolerate negativity."
0: "I prefer not to say."
0: "I have my own way of managing my class."

Response to classify:
Answer ID: {answer_id}
"{response}"

Provide only the classification (1 or 0) with no extra text.
"""
return prompt
```

A.3 Topic Coding By Response

As discussed earlier, to code responses the LLM is prompted to read each answer, then identify the most prominent themes (referred to as codes in qualitative research) in each answer, and to return up to two codes, with a code label of no more than three words, a code description, and a quote from the answer that embodies the idea behind the code.

We employ OpenAI's GPT-4o to obtain the codes, descriptions, and quotes for each response. This process is conducted on a per-instance basis, meaning that each response is evaluated through a separate API call.

The prompt for the "coding" is as follows:

```
def make_prompt(response):
    prompt = f"""
    This is an answer from a teacher to the question 'Describe your classroom
    management style'.

    Task:
    Identify the 2 most important themes in the response, provide a name for each
    theme in no more than 3 words, a 2-line meaningful and dense description of
    the theme, and extract a quote from the respondent no longer than 3 lines for
    each theme.

    Strict constraints:
    - Provide only two themes.
    - Do not use 'classroom management' as a name for a theme.
    - DO NOT include headers or additional text.
    - Format the response as a CSV file including only name, description, and
    quote.

    Input Text: "{response}"
    """
    return prompt
```

A.4 LLM-based Code Generation (HLTM Step 2)

Figure A.1 shows the distribution of counts of unique codes. Table A.3 shows the top 100 unique codes and their respective counts. Two points are worth noting. First, the distribution of code counts exhibits a pronounced concentration at the upper end of the spectrum. The most frequent code, "Positive Reinforcement," appears 522 times (5.9 percent) in the dataset, followed by the second most common code, "Clear Expectations," occurring 396 times (4.5 percent). The top five codes (which also include "Authoritative Approach" (291, 3.3 percent), "Structured Environment" (237, 2.7 percent), and "Relationship Building" (206, 2.3 percent)) together account for 18.55 percent of all entries in the dataset. The top 20 codes, each appearing at least 50 times, constitute 33.03 percent of all entries, and just

60–70 codes (approximately 5–6 percent of all unique codes) account for more than half of all occurrences in the dataset.

Second, these statistics are only based on exact code names and not semantic meaning. It is clear from inspecting the top few codes that, many of the most frequent codes will group together into broader themes. For instance, the second, third and fourth most frequent codes (Clear Expectations, Authoritative Approach and Structured Environment) are all concepts closely associated with Authoritative classroom management.

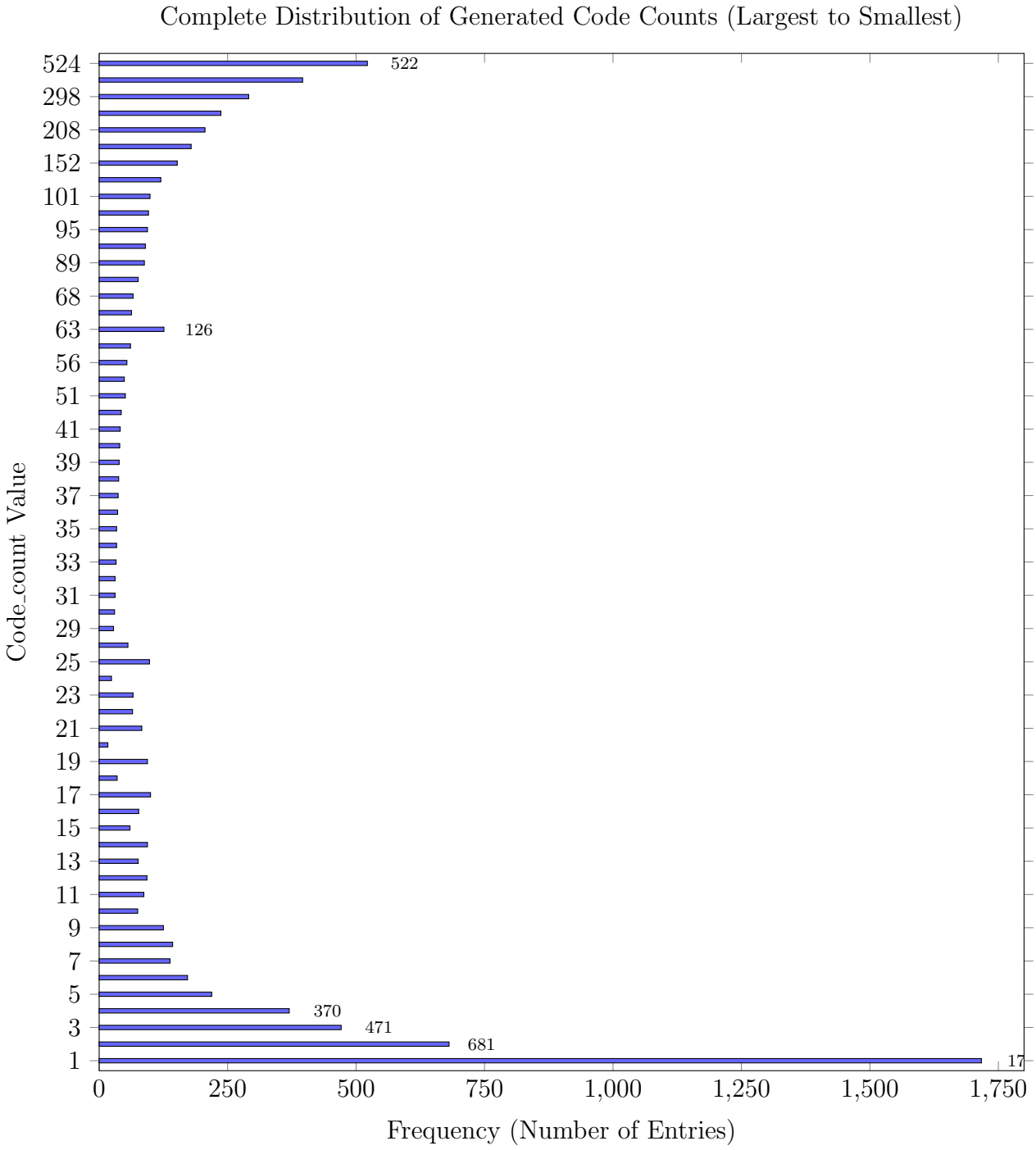


Figure A.1: Complete distribution of generated code counts

Table A.3: Code Label Counts for Top 50 Codes

Code Label	Count	Code Label	Count
Positive Reinforcement	522	Clear Communication	21
Clear Expectations	396	Engaging Lessons	21
Authoritative Approach	291	Structured Routines	21
Structured Environment	237	Modeling Behavior	20
Relationship Building	206	Interactive Engagement	19
Supportive Environment	179	Trust Building	19
Student Engagement	152	Reward System	19
Student Empowerment	120	Personal Connection	19
High Expectations	99	Authoritative Engagement	18
Respectful Environment	96	Building Rapport	18
Proactive Engagement	94	Proactive Environment	17
Positive Environment	90	Structure and Routine	17
Collaborative Learning	88	Structured Routine	17
Student-Centered Approach	76	Active Learning	17
Mutual Respect	66	Balanced Approach	17
Student Involvement	63	Structured Engagement	17
Positive Relationships	63	Respectful Interaction	16
Engagement Strategies	63	Routine Establishment	16
Collaborative Environment	61	Consistency in Expectations	16
Community Building	54	Fairness and Consistency	16
Active Engagement	51	Safe Environment	16
Proactive Approach	49	Individual Attention	15
Building Relationships	43	High Standards	15
Engagement Focused	41	Welcoming Environment	15
Engagement Focus	40	Supportive Learning Environment	15
Open Communication	39	Collaborative Rule Setting	15
Expectation Setting	38	Engaging Learning	14
Structured Learning	37	Encouraging Participation	14
Respectful Engagement	36	Consistent Expectations	14
Engaging Learning Environment	34	Authoritarian Approach	14
Student-Centered Learning	34	Routine and Structure	14
Interactive Learning	33	Positive Discipline	14
Parental Involvement	31	Firm and Fair	13
Balanced Authority	31	Behavior Modeling	13
Individualized Support	30	Student Ownership	13
Individualized Approach	28	Supportive Authority	13
Safe Learning Environment	28	Collaborative Rule-Making	13
Engaging Environment	28	Adaptive Teaching	12
Student-Centric Approach	25	Respectful Authority	12
Student Autonomy	25	Cooperative Learning	12
Engagement Techniques	25	Individual Support	12
Positive Engagement	24	Community Engagement	12
Firm but Fair	23	Student Responsibility	12
Adaptability in Teaching	23	Collaborative Expectations	12
Active Participation	22	Relaxed Environment	12
Engaged Learning	22	Inclusive Environment	11
Structured Approach	22	Engagement and Fun	11
Flexible Approach	21	Team Collaboration	11
Positive Learning Environment	21	Consistency in Rules	11
Clear Communication	21	Collaborative Approach	11

A.5 Codes Grouping Using BERTopic

We perform another check on this step to ensure that the grouping of codes into broader clusters is sensible. We inspect which codes were assigned to each cluster. The authors do this manually for all 35 initial clusters. As an example, the list of codes assigned to each of the top four clusters, with the underlying unique code count in brackets and total documents contributing to the cluster, is presented in Table A.4. Here, all codes assigned to each of the top four clusters seem appropriate, with the overwhelming majority of them very closely related to the broader cluster.

Table A.4: Top Four Clusters with the List of Associated Codes

(ID) Cluster Name	**Associated Codes (Frequency)**
(0) Student Engagement and Active Participation	Student Engagement (221) Engagement Focus (98) Engagement Strategies (88) Active Engagement (48) Engaging Environment (29) Interactive Learning (24) Fun Learning Environment (22) Engagement Techniques (21) Engaging Lessons (21) Engaging Learning (14) Engaged Learning (11) Active Participation (11) Engagement and Fun (9) Positive Engagement (8) Proactive Engagement (7) **Total Frequency: 983**
(1) Positive Reinforcement in Education	Positive Reinforcement (511) Positive Discipline (23) Reward System (23) Reward Systems (11) Intrinsic Motivation (9) Positive Behavior Support (8) Positive Behavior Focus (5) Encouragement Focus (5) Motivation Techniques (5) Motivational Techniques (4) Incentive-Based Rewards (4) Motivation Strategies (4) Behavior Incentives (4) Student Motivation (4) Token Economy (3) **Total Frequency: 788**
(2) Establishing Clear Classroom Expectations	Clear Expectations (301) Expectation Setting (25) Rules and Expectations (22) Expectations Setting (19) Rule Establishment (17) Behavior Expectations (17) Clear Communication (12) Behavioral Expectations (12) Firm Expectations (7) Rule Setting (7) Setting Expectations (6) Expectations and Rules (6) Expectations Clarity (6) Rules and Structure (5) Expectation Clarity (5) **Total Frequency: 651**

Continued on next page

Table A.4: Top Four Clusters with the List of Associated Codes (continued)

(ID) Cluster Name	Associated Codes (Frequency)
(3) Building Strong Relationships with Students	Relationship Building (167) Positive Relationships (66) Building Relationships (48) Building Rapport (23) Student Relationships (20) Rapport Building (16) Trust Building (13) Trust and Relationships (9) Personal Connection (7) Connection Building (5) Relational Approach (5) Positive Rapport (5) Student Connection (5) Trusting Relationships (5) Building Trust (5) **Total Frequency: 523**

Note: (ID) Cluster Name refers to the name and ID for each cluster, as reported in Table 3. Associated Codes (Frequency) shows the 15 most frequent codes per cluster.

A.6 Topic Classification

For this section of the analysis, we employ OpenAI's GPT-4o via API to generate indices representing the alignment of each response with the identified topics. The prompt instructs the model to analyze the response in the context of the predefined topics and guidelines, and then allocate 100 points across the topics based on prevalence.

This process is conducted on a per-instance basis, meaning that each response is evaluated through a separate API call.

The prompt for the classification is as follows:

```
def make_prompt(response):
    """
    Creates a prompt for GPT to classify a teacher's response into classroom
    management types.
    """

    prompt = f"""
    You are an assistant analyzing teacher responses to the question:
    **"Describe your classroom management style."**

    **Task:**
    1. Read the teacher's response carefully.
    2. Assign probability weights (in decimal form) to all of the following
    Classroom Management categories (listed below), based on their presence in
```

the response (e.g., 0.35 means 35%).
3. If multiple categories are present, distribute the probabilities accordingly.
4. **The sum of all probability weights must be exactly 1.00.**
5. Some key concepts may be expressed **without exact keyword matches**. Use reasoning beyond direct word matching.
6. Output only the classification results in the specified format. Do not add any other text.

Classroom Management Categories:

1. Authoritative_Approach_for_Structure_and_Accountability
Structured learning environments with explicit rules and expectations, including firm but fair consequences.
Teacher sets non-negotiable routines during the first classes, yet may invite limited but relevant student input when creating rules.
Emphasizes accountability, disciplinary actions or plans, corrective measures at the group or student level, and consistently high standards.
Keywords: 'structured learning', 'rules', 'procedures', 'expectations', 'discipline', 'consequences', 'high standards', 'love and logic', 'facts', 'goal oriented', 'authority', 'routines'

2. Student_Centered_Engagement_for_Empowering_Relationships
Teachers design lively, inclusive lessons that invite every learner to participatethrough debates, hands-on projects, and other group or class-wide activities that elevate student voice.
They flexibly tailor support to individual interests and emotional needs, building trust and a sense of belonging while encouraging autonomy and shared responsibility.
They generate respectful relationships through open communication and empathy, creating nurturing learning environments where every learner feels included and students feel valued.
Keywords: 'engagement', 'empowerment', 'trust', relationship, 'emotional support', learning environment, 'student voice', 'student-led', 'ownership', 'individual interests', 'non-verbal signals', 'flexibility', 'group activity', 'shared responsability', 'conferences', 'safe space', 'belonging'

3. Positive_Reinforcement_Strategies_and_Systems

```
Reinforces desired behaviors with praise, points, token economies, and other
tangible rewards; feedback is immediate and motivational.
Emphasizes prevention rather than punishment and rarely mentions consequences.
**Keywords**: 'positive feedback', 'token economy', 'point system', 'class
dojo', 'stickers', reward, 'praise', 'PBIS', behavior reinforcement,
'redirect', 'encourage', 'reflect on behavior'

**4. Modeling_Positive_Behavior_by_Example**
Teachers intentionally model the attitudes and behaviors they expect from
their students.
Rather than relying on rules or rewards, they lead by example, demonstrating
these qualities through their own actions.
Teachers often emphasize their own personal conduct and attitudes as a tool
to reinforce the desired norms.
**Keywords**: 'modeling behavior', 'role model', 'leading by example',
'emulate behavior'

- Make sure the sum of probabilities is exactly 1.00.
- Use two decimals for each probability (e.g., 0.35).
- Do not include any explanation, just the final lines with the categories
and probabilities.

**Example Response:**
"I maintain a structured classroom with clear rules, but I also reward good
behavior with praise and tokens. Students collaborate in small groups often."

**Example output with format (this is how a correct answer looks, strictly
follow this format, with two decimals and no extra commentary):**
- Authoritative_Approach_for_Structure_and_Accountability: 0.40
- Student_Centered_Engagement_for_Empowering_Relationships: 0.60
- Positive_Reinforcement_Strategies_and_Systems: 0.00
- Modeling_Positive_Behavior_by_Example: 0.00
"""

**Teacher Response:**
"{response}"
```

```
    return prompt
```

A.7 Inter-Rater Reliability: Additional Measures

This appendix reports additional inter-rater agreement measures for the same validation exercise described in the main text. We compute agreement on two complementary representations of the classifications. First, using the ordinal rankings assigned by each rater, we calculate rank-based agreement measures, including Kendall's τ, Spearman's ρ, and per-topic Cohen's κ. These measures focus on agreement in the relative ordering of topics. Second, we map each rater's rankings into normalized rank-implied topic-weight distributions that sum to one and compute distributional similarity and distance measures. Cosine similarity, Pearson's r, and Spearman's ρ capture similarity between the resulting distributions, while $\sqrt{\text{JSD}}$, Hellinger distance, Wasserstein$_1$, and Brier score capture distributional distance. Higher values of the similarity and correlation measures indicate greater agreement; lower values of the distance measures indicate greater agreement.

Table A.5: Additional Inter-Rater Reliability Measures for the Six-Topic Classification

Metric	**GPT vs H1**	**GPT vs H2**	**H1 vs H2**
Overall Rank Correlation			
Kendall τ	0.657	0.684	0.630
Spearman ρ	0.700	0.730	0.668
Per-Topic Cohen κ (Ranks)			
Authoritative Approach for Structure and Accountability	0.736	0.812	0.710
Student Engagement and Empowerment	0.643	0.528	0.506
Student-Centred Adaptive Teaching Approaches	0.536	0.447	0.335
Building Strong Relationships	0.558	0.681	0.502
Positive Reinforcement Strategies and Systems	0.725	0.716	0.781
Modelling Positive Behaviour by Example	0.345	0.506	0.643
Rank-Implied Full-Distribution Metrics			
Mean Similarity (1 = Perfect Match)			
Cosine	0.798	0.835	0.776
Pearson r	0.678	0.717	0.639
Spearman ρ	0.658	0.715	0.628

Continued on next page

Table A.5: Additional Inter-Rater Reliability Measures for the Six-Topic Classification (continued)

Metric	**GPT vs H1**	**GPT vs H2**	**H1 vs H2**
Mean Distance (0 = Perfect Match)			
$\sqrt{\text{JSD}}$	0.373	0.320	0.406
Hellinger	0.364	0.313	0.399
Wasserstein_1	0.071	0.046	0.085
Brier	0.043	0.029	0.047
Sample	195	212	195

Note: GPT = GPT-4o-mini. Human raters: H1 = principal investigator; H2 = PhD student. Reliability was computed on the same random sample of responses. Full-Distribution Metrics were computed on the full probability distribution of topics weights for each response. $\sqrt{\text{JSD}}$ is the square root of Jensen–Shannon divergence (base 2; bounded 0-1). Wasserstein_1 denotes the 1-Wasserstein (earth-mover) distance on the ordered theme axis. Higher cosine similarity, κ, and correlations, and lower values on the distance measures (JSD, Hellinger, Wasserstein1, Brier) indicate greater agreement.

Table A.6: Additional Inter-Rater Reliability Measures for the Four-Topic Classification

Metric	**GPT vs H1**	**GPT vs H2**	**H1 vs H2**
Overall Rank Correlation			
Kendall τ	0.783	0.802	0.777
Spearman ρ	0.817	0.836	0.817
Per-Topic Cohen κ (Ranks)			
(1.1) Authoritative Approach for Structure and Accountability	0.749	0.834	0.742
(1.2) Student-Centred Engagement for Empowering Relationships	0.668	0.603	0.608
(1.3) Positive Reinforcement Strategies and Systems	0.835	0.797	0.788
(1.4) Modelling Positive Behaviour by Example	0.605	0.621	0.679
Rank-Implied Full-Distribution Metrics			
Mean Similarity (1 = Perfect Match)			
Cosine	0.889	0.909	0.892
Pearson r	0.875	0.862	0.849
Spearman ρ	0.840	0.852	0.817
Mean Distance (0 = Perfect Match)			
$\sqrt{\text{JSD}}$	0.220	0.191	0.240
Hellinger	0.213	0.185	0.231

Continued on next page

Table A.6: Additional Inter-Rater Reliability Measures for the Four-Topic Classification (continued)

Metric	GPT vs H1	GPT vs H2	H1 vs H2
Wasserstein_1	0.067	0.050	0.078
Brier	0.044	0.032	0.040
Sample	202	211	201

Note: GPT = GPT-4o-mini. Human raters: H1 = principal investigator; H2 = Ph.D. student. Reliability was computed on the same random sample of responses. Full-Distribution Metrics were computed on the full probability distribution of topics weights for each response. $\sqrt{\text{JSD}}$ is the square root of Jensen–Shannon divergence (base 2; bounded 0-1). Wasserstein_1 denotes the 1-Wasserstein (earth-mover) distance on the ordered theme axis. Higher cosine similarity, κ, and correlations, and lower values on the distance measures (JSD, Hellinger, Wasserstein1, Brier) indicate greater agreement.

A.8 Prevalence of Topics by Question

Table A.7: Prevalence of Topics by Question

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
Question classroom management: *Describe your classroom management style. Why have you found this style to be effective?*			
(classroom management.1) Authoritative Approach for Structure and Accountability	Structured learning environments with explicit rules and expectations, including firm but fair consequences. Teacher sets non-negotiable routines during the first classes, yet may invite limited but relevant student input when creating rules. Emphasizes accountability, disciplinary actions or plans, corrective measures at the group or student level, and consistently high standards.	45%	0.35

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(classroom management.2) Student-Centered Engagement for Empowering Relationships	Teachers design lively, inclusive lessons that invite every learner to participate—through debates, hands-on projects, and other group or class-wide activities that elevate student voice. They flexibly tailor support to individual interests and emotional needs, building trust and a sense of belonging while encouraging autonomy and shared responsibility. They generate respectful relationships through open communication and empathy, creating nurturing learning environments where every learner feels included and valued.	45%	0.42
(classroom management.3) Positive Reinforcement Strategies and Systems	Reinforces desired behaviors with praise, points, token economies, and other tangible rewards; feedback is immediate and motivational. Emphasizes prevention rather than punishment and rarely mentions consequences.	7%	0.09
(classroom management.4) Modelling Positive Behaviour by Example	Teachers intentionally model the attitudes and behaviors they expect from their students. Rather than relying on rules or rewards, they lead by example, demonstrating these qualities through their own actions. Teachers often emphasize their own personal conduct and attitudes as a tool to reinforce the desired norms.	2%	0.11
Question students feel valued: *How do you build a classroom culture in which all students feel valued?*			
(students feel valued.1) High Expectations and Predictable Structures for Student Success	Teachers communicate consistent expectations and establish orderly routines that guide student behavior and academic engagement, fostering fairness and belonging. They emphasize accountability, structure, and a belief in students' ability and potential to meet high standards. These responses often describe how predictability helps students feel secure, motivated, focused, and able to succeed in a well-managed environment.	5%	0.07

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(students feel valued.2) Modeling Respectful Behavior by Example	Teachers highlight the importance of modeling the respectful, caring, and inclusive behavior they expect from students. They lead by example through their own actions and attitudes, such as greeting students warmly, listening actively, and showing kindness. These responses often emphasize intentional role modeling of behavior as a way to build mutual respect and reinforce positive norms.	4%	0.10
(students feel valued.3) Relational and Engaging Practices for Inclusive Classrooms	Teachers build trust and meaningful relationships through open communication, active listening, shared decision-making, and mutual respect. They create inclusive communities where students feel emotionally safe, known, valued, and actively involved. These responses highlight how engagement, empowerment, participation, and strong relationships drive motivation and classroom cohesion.	57%	0.36
(students feel valued.4) Personalized and Differentiated Student-Centered Instruction	Educators tailor instruction to students' individual learning needs, preferences, and strengths. They create flexible, student-centered environments that offer choices, student-led activities, and varied teaching methods. These responses emphasize how differentiated, personalized instruction helps students thrive through autonomy and responsiveness to their learning paths.	9%	0.12
(students feel valued.5) Culturally Inclusive Practices that Affirm Diversity and Identity	These responses highlight teaching practices that acknowledge, honor, and integrate students' cultural backgrounds and identities into learning. Teachers describe using inclusive curricula, representation in materials, and affirming language to help students feel seen and respected. The focus is on building cultural awareness through the appreciation and celebration of diversity to enrich classroom dialogue and promote representation.	12%	0.14

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(students feel valued.6) Positive Reinforcement and Recognition of Student Growth	Teachers focus on publicly recognizing effort, improvement, achievement, and individual contributions to encourage continued learning and confidence. They describe using positive reinforcement, systematic praise, rewards, incentives, and growth-oriented feedback to motivate students and support continual improvement. These responses value progress over perfection and frame mistakes as learning opportunities to reinforce a growth mindset and self-worth development.	14%	0.22
Question family communication: *What do you feel is the most effective way to communicate with families? Describe how you have used this/these technique(s).*			
(family communication.1) Building Trusting Relationships	Focuses on establishing strong relationships with families by fostering trust and collaboration through open communication, active listening, and empathy. It values parents' insights and encourages a supportive dialogue environment to enhance student success.	29%	0.23
(family communication.2) Active Engagement Through Regular and Direct Updates	Highlights the importance of engaging parents as partners through regular updates and consistent communication. It emphasizes direct interactions, such as phone calls, emails, newsletters, notes, and similar outreach to strengthen home-school connections and involve families in educational processes.	41%	0.28
(family communication.3) Positive Feedback and Reinforcement for Family Involvement	Centers on positive reinforcement and sharing student achievements with families. It involves parents in student growth with positive and constructive feedback.	1%	0.06
(family communication.4) Face-to-Face Interaction	Underscores the significance of face-to-face communication in building understanding between families and educators. It highlights the role of in-person participation and meetings in enhancing collaboration and effective communication.	9%	0.13

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(family communication.5) Tailored Communication Based on Family Preferences and Contexts	Emphasizes personalized communication with families, feeling comfortable with using various methods like emails, phone calls, and in-person meetings to meet individual needs. It highlights tools and translation services for inclusivity and effective participation and involvement across language barriers.	14%	0.23
(family communication.6) Digital Tools to Enhance Family-School Connection	Focuses on integrating technology for effective parent communication, highlighting tools like Class Dojo, Remind, and Seesaw for real-time updates. It also emphasizes using social media platforms to engage families and showcase school activities.	5%	0.07
Question student progress tracking: *How do you track student progress in your classroom?*			
(student progress tracking.1) Data-Driven Assessment and Instruction	This topic describes how teachers interpret student performance data—such as quiz scores, test results, or analytics reports—to make informed instructional decisions. These responses emphasize identifying patterns, tracking progress over time, or comparing results across students or groups. The goal is to adjust future lessons, target support, or refine teaching methods based on the insights gained from student outcomes.	26%	0.22
(student progress tracking.2) Baseline Assessment and Initial Diagnostics	This topic focuses on early instructional practices that identify what students know before teaching begins. Teachers describe using pre-assessment tests, entry quizzes, diagnostic activities, or informal observations to assess prior knowledge or readiness levels. These checkpoints help set goals or plan instruction, but occur before content is formally taught.	3%	0.05

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(student progress tracking.3) Personalized and Student-Centered Progress Monitoring	This topic involves adjusting teaching and monitoring to meet individual student needs and strengths. Teachers describe using one-on-one check-ins, custom learning plans, or individualized feedback to follow student progress. They emphasize tailoring instruction, offering targeted support, and tracking learning in flexible ways that respond to how each student develops over time.	10%	0.11
(student progress tracking.4) Student Engagement, Empowerment, and Holistic Development	This topic refers to how teachers track student participation, emotional growth, motivation, self-regulation, and sense of ownership in learning. It includes monitoring engagement, encouraging student-led goal setting, and documenting behaviors like collaboration, effort, or autonomy. Teachers also mention peer relationships, classroom behavior, and involvement of families as signs of broader development.	5%	0.09
(student progress tracking.5) Instructional Assessment Strategies and Feedback Practices	This topic centers on how teachers evaluate student understanding during or after instruction using specific assessment and evaluation techniques. Teachers describe using informal and formal evaluations, including formative tools like exit tickets, class questioning, or quizzes, and summative tasks like tests or projects. They also highlight how they give timely feedback to help students reflect or improve.	44%	0.35
(student progress tracking.6) Progress Tracking Tools and Documentation Systems	This topic focuses on the physical, visual, or digital systems teachers use to log, store, organize, and display student progress. These tools are often used for tracking rather than assessment and include spreadsheets, gradebooks, color-coded charts, dashboards, or portfolios, emphasizing technology integration. Teachers describe updating trackers or documenting progress over time without focusing on evaluation design.	12%	0.18

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
Question candidate qualities: *What qualities do you bring to this position which would make you the best candidate for the job?*			
(candidate qualities.1) Extensive Experience and Strong Educational Credentials	These responses emphasize substantial teaching experience combined with strong academic qualifications. Educators often mention leadership roles, subject-area expertise, or certifications in areas like special education or bilingual education. The focus is on a solid, well-rounded background that prepares them to handle diverse educational challenges.	29%	0.23
(candidate qualities.2) Cultural Awareness and Commitment to Inclusive Diversity	Teachers in this group stress the importance of inclusivity for students from diverse cultural, linguistic, and socioeconomic backgrounds. They often reference bilingualism and describe efforts to make all students feel respected and engaged. Culturally responsive, competent, and sensitive teaching combined with equity-driven practices are central.	6%	0.07
(candidate qualities.3) Relational Strengths Across Students Families and Colleagues	These responses focus on strong interpersonal skills and the ability to build trust with students, families, and colleagues, using storytelling and leveraging interpersonal skills to motivate and create meaningful, caring relationships. Teachers emphasize empathy, communication, and commitment to community engagement. Relationships are described as essential to supporting learning and collaboration.	20%	0.21
(candidate qualities.4) Collaborative Professionalism and Commitment to Growth	This theme highlights collaboration, reliability, and ongoing professional development. Teachers describe themselves as organized, reflective, and committed to teamwork and continuous improvement through mentorship and peer collaboration. Traits like strong work ethic, integrity, multitasking, and proactive problem-solving in fast-paced settings are presented as core strengths.	21%	0.21

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(candidate qualities.5) Student Centered Positive Engagement and Passion for Teaching	Educators in this theme express high enthusiasm for teaching and focus on keeping students engaged and motivated. They emphasize creating supportive, resilient, student-centered classrooms driven by energy and positivity. They also highlight adaptability, compassion, and collaboration. Passion for helping students learn and grow is central.	24%	0.29
Question classroom activities: *If I were to walk into your classroom what activities would I see you and the students doing?*			
(classroom activities.1) Active Engagement Through Creative and Physical Expression	Students are actively involved in energizing classroom environments that prioritize dynamic participation through movement, creativity, or artistic performance. Teachers describe learners "on their feet," engaging in music, drama, art, or physical education to support engagement and deepen understanding. These responses highlight enthusiasm and visible action, using expressive tools to make learning enjoyable and memorable.	17%	0.16
(classroom activities.2) Structured Skill-Building and Independent Learning	Students work within organized routines and clear expectations to build foundational academic, cognitive, or behavioral skills. Teachers describe structured centers, daily goals, modeling, or guided practice that help students develop reading, writing, vocabulary, math, or self-regulation. At the same time, students are expected to take ownership of their progress through independent work habits, self-directed tasks, quiet independent work, or following routines autonomously, including phrases like "reading centers" or "station-based learning."	16%	0.13

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(classroom activities.3) Collaborative Learning Through Discussion and Critical Thinking	Students collaborate in pairs, small groups, or whole-class conversations to analyze ideas, solve problems, and build shared understanding through dialogue. Classrooms are described as spaces where students discuss, debate, ask questions, and explain their reasoning during structured interactions. Teachers emphasize peer learning and academic discourse, often referencing Socratic seminars, think-pair-share, or meaningful conversations around content.	32%	0.25
(classroom activities.4) Interactive and Experiential Hands-On Learning Practices	Students engage with materials, tools, or real-world simulations to explore concepts through trial, error, and discovery. Teachers describe learners conducting experiments, building models, using manipulatives, or applying knowledge through project-based or tech-supported activities. The emphasis is on practical experience and hands-on exploration that promotes deep understanding through doing rather than observing.	19%	0.18
(classroom activities.5) Student-Centered and Differentiated Instruction	Instruction is tailored to each student's needs, interests, or learning pace, emphasizing personalization and flexibility in classroom design. Teachers describe adapting content, pacing, format, or strategies, using flexible grouping, choice-based tasks, scaffolds, or one-on-one support to foster individual progress. Classrooms often feature small groups, varied materials, and student-driven goals that reflect a responsive and inclusive approach.	6%	0.10
(classroom activities.6) Teacher Facilitation and Support	Teachers actively guide student learning by circulating, checking in, and offering timely feedback or clarification without directing every step. They monitor understanding, pose higher-order questions, and support students' thinking during independent or group work. This role is often described through phrases like "teacher is moving around," "guided practice," or "providing feedback while students work."	10%	0.19

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
Question active engagement: *What does active engagement look like in your classroom?*			
(active engagement.1) Peer and Group-Based Collaboration	Students actively engage in learning through interaction with peers or participation in group-based activities that require small-group discussion and teamwork. Teachers describe environments where collaboration enables deeper understanding, peer support, communication, and collective problem-solving in interactive settings. These responses often include references to working together on tasks, learning from each other, or building knowledge through joint effort.	38%	0.26
(active engagement.2) Teacher-Guided Engagement through Observation and Feedback	Teachers maintain student engagement and motivation by closely monitoring participation and offering responsive and constructive feedback based on observations and data tracking. They describe walking around the room, checking for understanding, or offering timely support as students work. These responses highlight a facilitative role where engagement stems from the teacher's presence, attention, and cues rather than direct instruction.	15%	0.14
(active engagement.3) Supportive Environment for Critical Thinking and Inquiry	Students engage when the classroom feels emotionally safe and intellectually open, allowing them to ask questions, express doubts, and think deeply. Teachers create a supportive and encouraging atmosphere for emotional engagement and critical thinking, emphasizing open communication, risk-taking, higher-order questioning, and problem-solving through inquiry-based learning. These environments promote active listening, empathy, exploration, and emotional engagement, with strong teacher-student relationships, and a nurturing, inclusive, judgment-free space.	9%	0.13

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(active engagement.4) Active and Experiential Participation	Students are physically or cognitively engaged through hands-on tasks, interactive games, real-world connections, or dynamic classroom activities. Teachers highlight the importance of movement, excitement, and interaction with materials or ideas to maintain attention and foster enthusiasm. These responses often describe project-based work, simulations, creative expression, physical involvement, and connecting learning to real-world experiences where students learn by doing.	27%	0.28
(active engagement.5) Empowered and Differentiated Student Learning	Teachers describe engagement emerging when students take ownership of their learning and when instruction adapts to individual strengths, needs, or preferences. These responses emphasize student-centered learning, responsibility, choice, independence, and leadership, alongside practices like personalized plans or flexible grouping. Engagement is framed as a result of students feeling capable, seen, and empowered in the learning process.	12%	0.18
Question difficult feedback: *What do you do to improve? Give an example of a time when you received difficult feedback from a peer/a student/a teacher/a supervisor. How did you respond in the moment? How did it affect your thinking? Your actions?*			
(difficult feedback.1) Self-Reflection for Growth Mindset	Teachers delve into self-reflection of their own practice by analyzing successes and missteps, welcoming constructive feedback, and setting specific improvement goals that reinforce a growth-oriented outlook. Narratives often mention looking back on lessons, revising beliefs about teaching effectiveness, and linking professional development plans to continuous self-improvement. These responses often reference evaluating past experiences, considering feedback thoughtfully, and making conscious adjustments to support continuous professional development.	35%	0.33

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(difficult feedback.2) Adaptive Teaching Methods	Teachers describe flexibly changing lesson structure, materials, or pacing in real time when student needs, assessment data, or peer feedback signal a better path forward. Accounts highlight experimenting with new approaches, tweaking strategies mid-lesson, and viewing adaptation as an iterative cycle that keeps instruction responsive and effective. These responses often highlight proactive changes that arise from observing students, analyzing performance, and integrating feedback into practice.	36%	0.26
(difficult feedback.3) Collaborative Communication and Relationship Building	Teachers emphasize building rapport through open communication, active listening, empathy, and teamwork with colleagues, students, and families to create a supportive classroom climate. Examples feature co-planning, peer coaching, empathetic conversations that defuse conflict, and intentional efforts to foster trust and mutual respect in everyday interactions. These responses also highlight teamwork, open dialogue, and building trust as essential practices for sustaining positive interactions in educational settings.	25%	0.25
(difficult feedback.4) Student-Centered Engagement Strategies	Teachers design interactive student-centered learning experiences that let students lead activities that give learners voice, choice, and hands-on participation, aiming to spark curiosity and sustain motivation. Descriptions include inquiry projects, differentiated tasks aligned with student interests, and structures that promote autonomy while keeping engagement high throughout the lesson arc. They highlight designing lessons that prioritize student needs, encourage interaction, and foster meaningful engagement with content.	4%	0.16
Question accommodate learning styles: *How do you accommodate different learning styles in your classroom?*			

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(accommodate learning styles.1) Multisensory Modality	Responses describe instruction aligned to sensory channels—visual, auditory, and kinesthetic/-tactile—to make concepts concrete and accessible through seeing, hearing, and doing. Teachers reference graphic organizers, anchor charts, demonstrations, videos, read-alouds, manipulatives, movement, or lab-style tasks that let students touch, build, or act out ideas for experiential learning. Language emphasizes sensory experiences as parallel pathways to comprehension.	22%	0.17
(accommodate learning styles.2) Active Engagement Tactics	Responses foreground activity formats that keep students participating throughout the lesson. Teachers mention stations or rotations, games and quick competitions, questioning routines, short discussions, warm-ups and brain breaks, or interactive tech that prompts frequent responses. The tone stresses momentum, attention, and student talk as built into the lesson flow.	5%	0.08
(accommodate learning styles.3) Collaborative Relationships and Empowering Climate	Responses highlight a supportive environment built on collaborative relationships, trust, and belonging, where students work together and feel known. Teachers describe routines that cultivate rapport and peer support alongside opportunities for student choice and voice in tasks, products, or norms. The emphasis falls on community, dignity, and shared ownership that invites participation, autonomy, and confidence.	10%	0.12
(accommodate learning styles.4) Individualized Support	Responses describe tailoring content, process, products, or pace to an individual learner's needs and interests. Teachers cite one-to-one scaffolds, personalized goals, conferences, accommodations or modifications, and targeted interventions based on the student's profile. The writing often references meeting each student where they are and adjusting supports as progress emerges.	37%	0.30

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(accommodate learning styles.5) Differentiated and Data Driven Methods	Responses explain class- or group-level differentiation using flexible grouping, tiered assignments, leveled materials, and scaffolded pathways informed by checks for understanding and progress data. Teachers reference formative assessment, exit tickets, rubrics, and data trackers to adjust instruction, alongside Universal Design for Learning choices and culturally relevant content to ensure access for diverse learners. The tone emphasizes planning structures that anticipate variability and use evidence to adapt whole-group instruction.	26%	0.33
Question philosophy learning growth: *Describe your philosophy of how children learn and grow.*			
(philosophy learning growth.1) Personalized and Differentiated Instruction for Holistic Development	Teachers describe tailoring learning to each student's needs, interests, strengths, and pace, using flexible grouping, accommodations, and varied pathways to mastery. They frame growth across cognitive, social-emotional, and physical domains as holistic development, emphasizing inclusive access and culturally responsive adaptations so every learner can progress. Evidence often names concrete plans or adjustments—differentiated tasks, alternative materials, varied assessments, or pacing supports—that align instruction with the whole learner.	45%	0.33
(philosophy learning growth.2) Trusting Relationships and Supportive Classroom Climate	Teachers emphasize a caring, respectful atmosphere where students feel safe, known, and valued, highlighting rapport, empathy, and consistent, positive interactions. They describe routines and community practices that build belonging and social skills, including peer support and collaboration framed as mutual care and inclusion. Descriptions frequently mention trust, connection, and warmth as the day-to-day conditions that enable students to participate and grow.	34%	0.35

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(philosophy learning growth.3) Hands On Learning and Role Modeling for Growth Mindset	Teachers center learning by doing—projects, experiments, labs, manipulatives, and making—while guiding practice through demonstrations, think-alouds, and "I do–We do–You do" routines. They cultivate student agency and ownership, pairing clear models of expert behavior with feedback that praises effort, persistence, and learning from mistakes. Language often references hands-on experiences, role modeling, growth mindset, and positive reinforcement as the mechanisms that keep students engaged and progressing through challenging, authentic tasks.	21%	0.32
Question learning philosophy: *What is your philosophy on learning?*			
(learning philosophy.1) Personalized and Student-Centered Differentiation	These responses describe instruction tailored to individual learners through adjustable content, pacing, modalities, and assessments that align to unique needs, strengths, interests, and learning profiles. Teachers reference concrete mechanisms such as Universal Design for Learning, flexible grouping, accommodations, choice pathways, and multiple ways to demonstrate understanding. The focus remains on customizing learning experiences for each student rather than on classroom climate or general activity formats.	20%	0.24
(learning philosophy.2) Equitable and Inclusive Caring Classroom Community	These responses center on a welcoming, culturally and linguistically responsive environment where students feel safe, respected, and known, and where belonging and collaboration are daily practices. Teachers emphasize fairness, access, and relationship-building routines that nurture trust, empathy, and participation across identities and backgrounds. The emphasis is the community and climate that develop students' potential rather than individualized instructional tailoring or specific activity designs.	22%	0.25

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(learning philosophy.3) Growth Mindset and High Expectations	These responses frame learning as improvement through effort, challenge, and feedback, treating errors as information and emphasizing perseverance and productive struggle. Teachers communicate clear, challenging standards and a consistent belief that all students can meet them through revision, practice, and sustained commitment. Language emphasizes setting high expectations, a culture of achievement, rigor, resilience, and progress over perfection.	21%	0.16
(learning philosophy.4) Active and Experiential Learning with Student Agency for Lifelong Learning	These responses emphasize active participation through engagement, hands-on experiences, inquiry, problem- or project-based tasks, and real-world applications paired with student voice, choice, and ownership. Teachers act as facilitators while students set goals, make decisions, and reflect, cultivating curiosity, critical thinking, intrinsic motivation, and durable skills that extend beyond formal education. The focus is the activity design, student empowerment, ownership and independence, integrating holistic development that drive authentic, real-world and transferable learning over time.	37%	0.36
Question feedback sources: *To whom can you look for valuable feedback? How did you respond to that feedback?*			
(feedback sources.1) Mentorship and Supervisory Feedback for Professional Growth	Teachers describe structured feedback from mentors, supervisors, administrators, or instructional coaches that guides professional learning through observation notes, evaluation cycles, and targeted coaching. Narratives emphasize expert advice, reliable sources, and formal processes that refine instruction, document performance, and support advancement. Concrete references to walkthroughs, post-observation conferences, and actionable suggestions signal this theme.	39%	0.28

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(feedback sources.2) Supportive and Trust-Based Feedback Relationships	Teachers portray feedback emerging within their families, friends, or ongoing professional relationships grounded in trust, psychological safety, and candid dialogue. Accounts highlight rapport, empathy, and openness that invite honest insights and reflective conversation outside formal evaluation structures. References to regular check-ins, safe spaces for sharing challenges, and constructive back-and-forth mark this theme.	8%	0.10
(feedback sources.3) Peer Collaboration and Collegial Feedback Exchange	Teachers recount lateral exchange with colleagues—co-planning, lesson study, peer observation, and shared reflection on practice—to generate specific, actionable feedback. Descriptions point to team meetings, PLCs, and strategy swaps that inform immediate instructional adjustments. Mentions of observing each other's classes, debriefs, and integrating peer suggestions identify this theme.	18%	0.17
(feedback sources.4) Multi-Source and Stakeholder Feedback	Teachers draw feedback from multiple perspectives—administrators, colleagues, students, and families—using tools like surveys, conferences, and communication logs to assemble a fuller picture of classroom practice. Narratives emphasize stakeholder engagement, community input, and feedback loops that synthesize diverse signals into instructional decisions. References to open dialogue with parents and multichannel input characterize this theme.	12%	0.11
(feedback sources.5) Student Engagement and Instructional Feedback	Teachers treat student responses—questions, participation, work products, quick checks, and direct comments—as immediate feedback on lesson clarity and pacing. Descriptions highlight reading the room, tracking engagement, and adapting explanations or tasks based on student signals. Mentions of exit tickets, formative checks, and student voice indicate this theme.	7%	0.08

Continued on next page

Table A.7: Prevalence of Topics by Question (continued)

(ID) Theme	Theme Summary/Classification Instruction	Preval.	
		Dmnt Tpc	Avg Scr
(feedback sources.6) 6. Reflective Response to Feedback for Improvement	Teachers focus on how they process feedback—examining evidence, interpreting critique, and planning concrete changes to practice—then monitoring results. Narratives center on self-assessment, action steps, and iteration rather than the feedback source, referencing reflection notes, revisions, and follow-up implementation. Clear markers include identifying areas to refine and documenting adjustments over time.	16%	0.27

A.9 Dispersion of Theme Mixtures by Question—All Measures

This appendix reports additional diagnostics for the question-level topical distribution analysis in Table 9. The first table expands the dispersion measures by adding within-response diagnostics, including normalized Gini–Simpson evenness, Aitchison concentration, the prevalence of near-zero topic weights, and the prevalence of monothematic responses. These measures complement the main-text across-response dispersion measure by describing whether typical responses spread weight across several topics or concentrate heavily on one topic. The second table expands the selection-ratio diagnostic by reporting whether extreme racial selection ratios more often imply under- or over-representation relative to White applicants, which are our reference group.

Table A.8: Additional Dispersion Diagnostics for Topic Mixtures by Question

Question	Within Response		Between Response			n	K
	Gini-Simpson	Aitchison	ilr Total Var.	$< 1\%$	$\geq 80\%$		
classroom activities	0.77	20.45	365.76	98.57	4.42	6296	6
students feel valued	0.79	19.66	272.11	94.71	3.95	5896	6
learning philosophy	0.75	14.97	169.88	74.39	19.46	4770	4
classroom management	0.67	15.95	180.76	92.52	13.58	4629	4
student progress tracking	0.60	20.45	332.96	99.11	12.93	1910	6
philosophy learning growth	0.93	0.65	74.60	36.35	12.66	1622	3
candidate qualities	0.78	18.91	239.52	90.03	8.60	1465	5
family communication	0.77	20.45	315.69	99.55	6.05	1342	6
active engagement	0.78	18.91	288.85	95.70	12.21	1164	5
accommodate learning styles	0.78	18.93	227.96	89.84	8.32	817	5
difficult feedback	0.88	1.20	101.97	46.69	0.00	664	4
feedback sources	0.77	20.38	281.65	94.70	5.30	434	6

Notes: *Gini–Simpson* denotes normalized Gini–Simpson evenness, $(1 - \sum_i p_i^2)/(1 - 1/K) \in [0, 1]$; higher values indicate more even within-answer mixes across K themes. *Aitchison* distance is computed in clr space as $\|\mathrm{clr}(x)\|_2$, where $\mathrm{clr}(x)_i = \log x_i - \frac{1}{K}\sum_j \log x_j$; 0 equals a perfectly even mix, larger values indicate more concentrated answers. We report Aitchison and Gini-Simpson median (typical answer). *ilr Total Var.* is the trace of the covariance matrix of isometric log-ratio (ilr) coordinates across answers and reflects between-answer heterogeneity. % of near-zero ($< 1\%$) is the share of answers with any theme below 1%; % of monothematic ($\geq 80\%$) is the share with a single theme at least 80%. Zeros are replaced with a small ε and compositions are re-closed before log-ratio transforms.

Table A.9: Prevalence and Direction of Extreme Implied Selection Ratios by Question

Question	Share of Extreme Selection Ratios	Relative to White Applicants	
		Under	Over
classroom activities	0.29	0.08	0.14
students feel valued	0.21	0.06	0.19
learning philosophy	0.25	0.11	0.11
classroom management	0.22	0.04	0.13
student progress tracking	0.23	0.17	0.08
philosophy learning growth	0.46	0.22	0.28
candidate qualities	0.40	0.17	0.30
family communication	0.31	0.17	0.19
active engagement	0.33	0.17	0.20
accommodate learning styles	0.53	0.27	0.33
difficult feedback	0.41	0.21	0.33
feedback sources	0.33	0.22	0.19

Notes: The Share of Extreme Selection Ratios is the proportion of implied demographic selection ratios (Black/White, Hispanic/White, Asian/White, and Female/Male) that fall below 0.8 or above 1.25. Under- and over-representation are computed using racial selection ratios relative to White applicants only (Black-/White, Hispanic/White, and Asian/White), where Under denotes ratios below 0.8 and Over denotes ratios above 1.25. Implied hiring outcomes are generated by comparing applicants with the highest or lowest scores on each topic. For each question, there are $8K$ total implied ratios and $6K$ racial ratios relative to White applicants. These diagnostics are descriptive flags and do not represent actual hiring outcomes.

A.10 Estimated Relationships Between Question Topic Scores and Individual Demographic Characteristics

Table A.10: Estimated Relationship Between Classroom Management Question Topic Scores and Individual Characteristics (Base = classroom management.4)

Covariates	**classroom management.1**	**classroom management.2**	**classroom management.3**
Female (= 1)	0.798***	-0.152	1.372***
	(0.279)	(0.223)	(0.265)
Black (= 1)	1.048**	0.997***	0.259
	(0.439)	(0.352)	(0.417)
AI/PI (= 1)	-0.624	-1.572	-0.104
	(1.416)	(1.134)	(1.346)
Asian (= 1)	-0.947**	-0.177	-0.487
	(0.477)	(0.382)	(0.454)
Hispanic (= 1)	0.243	0.127	-0.105
	(0.488)	(0.391)	(0.464)
Mixed (= 1)	0.518	1.129**	0.0517
	(0.667)	(0.535)	(0.635)
PNS/Miss	0.332	0.155	1.299***
	(0.347)	(0.278)	(0.330)
Exp0 ($Exp = 0$)	-0.693**	-0.134	-0.554*
	(0.348)	(0.278)	(0.330)
Exp1 ($1 \geq Exp > 0$)	-0.558	-0.592	-0.137
	(0.564)	(0.452)	(0.536)
Exp2 ($2 \geq Exp > 1$)	-0.110	-0.587	-0.558
	(0.594)	(0.476)	(0.565)
Exp5 ($5 \geq Exp > 2$)	0.277	-0.0479	0.188
	(0.361)	(0.289)	(0.344)
Exp10 ($10 \geq Exp > 5$)	-0.482	-0.297	-0.319
	(0.329)	(0.263)	(0.313)
Constant	2.396***	4.488***	-3.155***
	(0.307)	(0.246)	(0.292)
Mean (log-ratio)	2.725	4.398	-2.312
N	4167	4167	4167

Table A.11: Estimated Relationship between Students Feel Valued Question Topic Scores and Individual Characteristics (Base = Students Feel Valued.6)

Covariates	**students feel valued.1**	**students feel valued.2**	**students feel valued.3**	**students feel valued.4**	**students feel valued.5**
Female (= 1)	0.164	0.447***	0.133	-0.156	0.778***
	(0.153)	(0.170)	(0.126)	(0.218)	(0.202)
Black (= 1)	0.536**	-0.552**	0.379*	0.994***	1.030***
	(0.245)	(0.275)	(0.204)	(0.365)	(0.327)
AI/PI (= 1)	0.000	0.000	-0.000***	-1.575	-0.000
	(0.000)	(0.000)	(0.000)	(1.323)	(0.000)
Asian (= 1)	0.000	-0.000	0.000***	-0.179	0.000***
	(0.000)	(0.000)	(0.000)	(0.365)	(0.000)
Hispanic (= 1)	-0.000**	-0.000***	0.000	0.124	-0.000
	(0.000)	(0.000)	(0.000)	(0.398)	(0.000)
Mixed (= 1)	-0.000*	0.000*	-0.000	1.126**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.465)	(0.000)
PNS/Miss	0.000*	-0.000**	0.000**	0.150	0.000**
	(0.000)	(0.000)	(0.000)	(0.278)	(0.000)
Exp0 ($Exp = 0$)	-0.562**	-0.719***	0.098	-0.137	0.646**
	(0.223)	(0.254)	(0.203)	(0.281)	(0.308)
Exp1 ($1 \geq Exp > 0$)	0.370	0.038	0.891***	-0.596	0.918*
	(0.369)	(0.383)	(0.276)	(0.467)	(0.473)
Exp2 ($2 \geq Exp > 1$)	-0.553	-0.160	0.650**	-0.591	0.772*
	(0.353)	(0.388)	(0.268)	(0.486)	(0.454)
Exp5 ($5 \geq Exp > 2$)	-0.285	0.153	0.353**	-0.051	0.264
	(0.220)	(0.242)	(0.171)	(0.291)	(0.286)
Exp10 ($10 \geq Exp > 5$)	-0.539***	-0.220	0.267*	-0.301	0.241
	(0.201)	(0.226)	(0.162)	(0.261)	(0.264)
Constant	-5.613***	-3.721***	0.933***	4.495***	-4.469***
	(0.153)	(0.173)	(0.132)	(0.238)	(0.202)
Mean (log-ratio)	-5.736	-3.674	1.271	-3.878	-3.623
N	5471	5471	5471	5471	5471

Table A.12: Estimated Relationship Between Family Communication Question Topic Scores and Individual Characteristics (Base = Family communication.6)

Covariates	**family communi-cation.1**	**family communi-cation.2**	**family communi-cation.3**	**family communi-cation.4**	**family communi-cation.5**
Female (= 1)	0.086	0.306	0.079	-0.691	0.014
	(0.467)	(0.397)	(0.414)	(0.455)	(0.444)
Black (= 1)	-0.811	-0.563	-0.605	-0.344	-0.445
	(0.548)	(0.454)	(0.465)	(0.515)	(0.528)
AI/PI (= 1)	-0.000*	-0.000	0.000***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	-0.000	-0.000***	-0.000***	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	-0.000	0.000**	0.000***	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	-0.000	0.000**	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	-0.000	0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	0.680	-1.125	-0.437	-0.912	-0.342
	(0.847)	(0.694)	(0.681)	(0.789)	(0.794)
Exp1 ($1 \geq Exp > 0$)	2.366***	2.494***	3.070***	0.308	0.688
	(0.896)	(0.697)	(0.765)	(0.849)	(0.896)
Exp2 ($2 \geq Exp > 1$)	1.058	0.273	2.073**	-0.989	-0.823
	(1.033)	(0.864)	(0.940)	(0.963)	(0.851)
Exp5 ($5 \geq Exp > 2$)	0.978	1.370***	1.096*	-0.792	0.981
	(0.635)	(0.524)	(0.575)	(0.623)	(0.661)
Exp10 ($10 \geq Exp > 5$)	1.711***	0.353	1.323**	-0.345	0.442
	(0.601)	(0.524)	(0.545)	(0.590)	(0.584)
Constant	3.557***	4.882***	-0.270	3.118***	5.319***
	(0.523)	(0.457)	(0.463)	(0.499)	(0.510)
Mean (log-ratio)	4.313	5.346	0.455	2.222	5.471
N	1088	1088	1088	1088	1088

Table A.13: Estimated Relationship Between Student Progress Tracking Question Topic Scores and Individual Characteristics (Base = Student Progress Tracking.6)

Covariates	**student progress tracking.1**	**student progress tracking.2**	**student progress tracking.3**	**student progress tracking.4**	**student progress tracking.5**
Female (= 1)	0.260	0.079	0.802**	0.158	0.774*
	(0.415)	(0.352)	(0.380)	(0.380)	(0.421)
Black (= 1)	-0.157	-0.368	-1.466***	-0.921**	-0.367
	(0.463)	(0.390)	(0.420)	(0.423)	(0.461)
AI/PI (= 1)	0.000**	0.000***	0.000**	0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	-0.000	-0.000	-0.000	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	-0.000	-0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	-0.000	-0.000	-0.000***	-0.000**	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	0.000	0.000	0.000***	0.000*	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	0.286	0.343	0.348	1.162**	0.256
	(0.634)	(0.517)	(0.569)	(0.568)	(0.639)
Exp1 ($1 \geq Exp > 0$)	0.947	0.946	0.259	1.076	1.256
	(0.820)	(0.741)	(0.764)	(0.759)	(0.862)
Exp2 ($2 \geq Exp > 1$)	1.570*	1.260*	1.220	1.832**	1.020
	(0.826)	(0.760)	(0.797)	(0.812)	(0.848)
Exp5 ($5 \geq Exp > 2$)	1.613***	0.777	1.332**	1.109**	0.312
	(0.571)	(0.494)	(0.552)	(0.543)	(0.592)
Exp10 ($10 \geq Exp > 5$)	0.842	0.721	0.748	0.892*	0.798
	(0.584)	(0.496)	(0.535)	(0.521)	(0.582)
Constant	0.313	-3.647***	-2.046***	-2.300***	2.143***
	(0.451)	(0.385)	(0.411)	(0.409)	(0.460)
Mean (log-ratio)	1.136	-3.185	-1.424	-1.655	2.888
N	1699	1699	1699	1699	1699

Table A.14: Estimated Relationship Between Candidate Qualities Question Topic Scores and Individual Characteristics (Base = Candidate Qualities.5)

Covariates	**candidate qualities.1**	**candidate qualities.2**	**candidate qualities.3**	**candidate qualities.4**
Female (= 1)	-0.694*	0.514	0.523	0.585
	(0.396)	(0.327)	(0.362)	(0.398)
Black (= 1)	0.041	-0.479	0.259	-0.033
	(0.771)	(0.555)	(0.718)	(0.774)
AI/PI (= 1)	-0.000**	0.000***	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	0.000***	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	0.000**	-0.000***	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	-0.000	-0.000***	-0.000*	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	-0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	-0.893	-0.502	-1.215**	-0.287
	(0.587)	(0.466)	(0.531)	(0.606)
Exp1 ($1 \geq Exp > 0$)	-2.770***	0.318	-0.126	-0.540
	(0.915)	(0.853)	(0.928)	(0.944)
Exp2 ($2 \geq Exp > 1$)	-2.362***	0.842	-0.523	-1.694**
	(0.845)	(0.755)	(0.711)	(0.806)
Exp5 ($5 \geq Exp > 2$)	-1.577***	0.217	-0.321	0.483
	(0.572)	(0.489)	(0.524)	(0.564)
Exp10 ($10 \geq Exp > 5$)	-1.781***	-0.026	-0.378	-0.822
	(0.530)	(0.446)	(0.488)	(0.538)
Constant	-1.028**	-6.429***	-2.077***	-2.339***
	(0.412)	(0.338)	(0.397)	(0.437)
Mean (log-ratio)	-2.533	-6.184	-2.177	-2.302
N	1238	1238	1238	1238

Table A.15: Estimated Relationship Between Classroom Activities Question Topic Scores and Individual Characteristics (Base = Classroom Activities.6)

Covariates	**classroom activities.1**	**classroom activities.2**	**classroom activities.3**	**classroom activities.4**	**classroom activities.5**
Female (= 1)	0.856***	1.557***	0.960***	1.597***	1.835***
	(0.202)	(0.197)	(0.183)	(0.212)	(0.215)
Black (= 1)	0.689*	-0.722*	-1.559***	-1.144***	-1.368***
	(0.391)	(0.375)	(0.347)	(0.425)	(0.395)
AI/PI (= 1)	-0.000	-0.000*	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	-0.000	-0.000**	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	0.000	-0.000	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	0.000	-0.000*	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	0.000*	-0.000**	-0.000***	-0.000**	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	0.943***	-0.955***	-0.755**	0.183	-0.476
	(0.329)	(0.304)	(0.299)	(0.348)	(0.336)
Exp1 ($1 \geq Exp > 0$)	-0.202	-0.954**	-0.194	-0.120	-0.021
	(0.428)	(0.408)	(0.374)	(0.454)	(0.444)
Exp2 ($2 \geq Exp > 1$)	-0.574	0.033	-0.006	0.268	0.107
	(0.409)	(0.421)	(0.388)	(0.436)	(0.468)
Exp5 ($5 \geq Exp > 2$)	0.373	0.121	0.501*	0.573*	0.706**
	(0.286)	(0.283)	(0.257)	(0.300)	(0.306)
Exp10 ($10 \geq Exp > 5$)	0.111	-0.194	0.284	0.195	0.200
	(0.276)	(0.274)	(0.248)	(0.292)	(0.298)
Constant	-2.683***	-3.261***	-0.207	-2.269***	-4.059***
	(0.212)	(0.201)	(0.192)	(0.224)	(0.224)
Mean (log-ratio)	-1.99	-2.683	0.253	-1.289	-3.033
N	5822	5822	5822	5822	5822

Table A.16: Estimated Relationship Between Active Engagement Question Topic Scores and Individual Characteristics (Base = Active Engagement.5)

Covariates	**active engagement.1**	**active engagement.2**	**active engagement.3**	**active engagement.4**
Female (= 1)	0.549	-0.354	-0.057	0.143
	(0.562)	(0.500)	(0.484)	(0.510)
Black (= 1)	-0.651	0.244	-0.138	0.296
	(0.607)	(0.545)	(0.532)	(0.544)
AI/PI (= 1)	-0.000	-0.000	-0.000*	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	-0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	-0.956	-0.703	-0.945	-0.135
	(0.776)	(0.676)	(0.648)	(0.695)
Exp1 ($1 \geq Exp > 0$)	0.461	-0.063	1.200	0.650
	(1.483)	(1.218)	(1.147)	(1.179)
Exp2 ($2 \geq Exp > 1$)	-0.660	0.864	-0.296	0.350
	(1.175)	(0.826)	(0.989)	(0.985)
Exp5 ($5 \geq Exp > 2$)	-0.313	0.567	0.699	-0.118
	(0.807)	(0.721)	(0.698)	(0.724)
Exp10 ($10 \geq Exp > 5$)	0.532	0.592	0.637	0.255
	(0.782)	(0.699)	(0.668)	(0.718)
Constant	1.082*	-1.294**	-1.069**	2.161***
	(0.591)	(0.511)	(0.492)	(0.519)
Mean (log-ratio)	0.971	-1.269	-1.064	2.374
N	1071	1071	1071	1071

Table A.17: Estimated Relationship Between Difficult Feedback Question Topic Scores and Individual Characteristics (Base = Difficult Feedback.4)

Covariates	**difficult feedback.1**	**difficult feedback.2**	**difficult feedback.3**
Female (= 1)	-0.756**	-0.789	-0.110
	(0.363)	(0.614)	(0.497)
Black (= 1)	0.212	0.780	0.094
	(0.510)	(0.791)	(0.679)
AI/PI (= 1)	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Asian (= 1)	0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)
Mixed (= 1)	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
PNS/Miss	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	-0.010	-0.119	0.243
	(0.465)	(0.792)	(0.692)
Exp1 ($1 \geq Exp > 0$)	0.659	0.486	1.049
	(0.998)	(1.651)	(1.337)
Exp2 ($2 \geq Exp > 1$)	0.127	0.603	0.771
	(0.784)	(1.298)	(0.991)
Exp5 ($5 \geq Exp > 2$)	-0.188	-0.291	0.625
	(0.473)	(0.855)	(0.685)
Exp10 ($10 \geq Exp > 5$)	0.543	0.860	1.197*
	(0.487)	(0.765)	(0.699)
Constant	2.488***	0.077	0.431
	(0.335)	(0.577)	(0.510)
Mean (log-ratio)	2.409	0.116	0.912
N	571	571	571

Table A.18: Estimated Relationship Between Philosophy Learning Growth Question Topic Scores and Individual Characteristics (Base = Philosophy Learning Growth.3)

Covariates	**philosophy learning growth.1**	**philosophy learning growth.2**
Female (= 1)	0.259	0.170
	(0.357)	(0.295)
Black (= 1)	-0.721	-0.190
	(0.504)	(0.414)
AI/PI (= 1)	-0.000	-0.000
	(0.000)	(0.000)
Asian (= 1)	-0.000	-0.000
	(0.000)	(0.000)
Hispanic (= 1)	0.000	-0.000
	(0.000)	(0.000)
Mixed (= 1)	0.000	0.000
	(0.000)	(0.000)
PNS/Miss	-0.000	-0.000
	(0.000)	(0.000)
Exp0 ($Exp = 0$)	0.334	0.735
	(0.705)	(0.594)
Exp1 ($1 \geq Exp > 0$)	-0.549	-0.147
	(0.708)	(0.614)
Exp2 ($2 \geq Exp > 1$)	-0.531	-0.363
	(0.629)	(0.496)
Exp5 ($5 \geq Exp > 2$)	-0.024	0.244
	(0.485)	(0.394)
Exp10 ($10 \geq Exp > 5$)	0.106	0.225
	(0.442)	(0.365)
Constant	-1.238***	-0.226
	(0.396)	(0.328)
Mean (log-ratio)	-1.237	-0.013
N	1333	1333

Table A.19: Estimated Relationship between Accommodate Learning Styles Question Topic Scores and Individual Characteristics (Base = Accommodate Learning Styles.5)

Covariates	**accommodate learning styles.1**	**accommodate learning styles.2**	**accommodate learning styles.3**	**accommodate learning styles.4**
Female (= 1)	1.012**	0.399	0.001	-0.434
	(0.464)	(0.454)	(0.507)	(0.492)
Black (= 1)	-0.457	-0.426	-0.619	0.547
	(0.579)	(0.571)	(0.612)	(0.623)
AI/PI (= 1)	0.000	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	0.000	0.000	0.000*	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	-0.000*	-0.000**	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	-0.000*	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	-1.071	0.052	0.120	-0.329
	(0.687)	(0.660)	(0.718)	(0.732)
Exp1 ($1 \geq Exp > 0$)	-1.460	-0.888	0.268	-0.476
	(1.099)	(1.037)	(1.102)	(1.139)
Exp2 ($2 \geq Exp > 1$)	-0.388	-1.109	0.226	-1.447
	(1.073)	(0.973)	(1.065)	(1.152)
Exp5 ($5 \geq Exp > 2$)	-0.139	0.139	0.731	0.263
	(0.677)	(0.664)	(0.710)	(0.709)
Exp10 ($10 \geq Exp > 5$)	0.043	0.230	0.134	-0.741
	(0.648)	(0.636)	(0.681)	(0.690)
Constant	-4.725***	-6.637***	-5.449***	-1.325**
	(0.526)	(0.516)	(0.565)	(0.603)
Mean (log-ratio)	-4.554	-6.491	5.354	-1.761
N	643	643	643	643

Table A.20: Estimated Relationship between Learning Philosophy Question Topic Scores and Individual Characteristics (Base = Learning Philosophy.4)

Covariates	**learning philosophy.1**	**learning philosophy.2**	**learning philosophy.3**
Female (= 1)	1.031***	1.403***	0.357*
	(0.207)	(0.192)	(0.187)
Black (= 1)	-0.302	0.159	0.470
	(0.352)	(0.330)	(0.323)
AI/PI (= 1)	0.000**	0.000	0.000**
	(0.000)	(0.000)	(0.000)
Asian (= 1)	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Mixed (= 1)	0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)
PNS/Miss	0.000	0.000	-0.000**
	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	-0.535*	-0.263	-1.058***
	(0.299)	(0.280)	(0.279)
Exp1 ($1 \geq Exp > 0$)	-1.080**	-0.605	-0.717*
	(0.494)	(0.465)	(0.420)
Exp2 ($2 \geq Exp > 1$)	-1.266***	-0.316	-1.283***
	(0.486)	(0.453)	(0.450)
Exp5 ($5 \geq Exp > 2$)	-0.062	0.006	-0.509*
	(0.302)	(0.278)	(0.272)
Exp10 ($10 \geq Exp > 5$)	-0.301	-0.180	-0.238
	(0.279)	(0.255)	(0.246)
Constant	-4.595***	-3.372***	-2.149***
	(0.208)	(0.196)	(0.190)
Mean (log-ratio)	-4.348	-2.724	-2.334
N	4331	4331	4331

Table A.21: Estimated Relationship between Feedback Sources Question Topic Scores and Individual Characteristics (Base = Feedback Sources.6)

Covariates	**feedback sources.1**	**feedback sources.2**	**feedback sources.3**	**feedback sources.4**	**feedback sources.5**
Female (= 1)	-0.106	0.510	0.504	-0.342	0.285
	(0.716)	(0.627)	(0.766)	(0.709)	(0.687)
Black (= 1)	-0.470	-0.312	-1.110	-0.409	0.020
	(0.726)	(0.642)	(0.765)	(0.728)	(0.664)
AI/PI (= 1)	-0.000	0.000	-0.000	-0.000**	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Asian (= 1)	0.000	0.000***	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Hispanic (= 1)	-0.000	-0.000*	-0.000	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mixed (= 1)	0.000	0.000	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PNS/Miss	-0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Exp0 ($Exp = 0$)	-0.213	0.067	-0.442	-1.847*	0.616
	(0.942)	(0.839)	(1.009)	(1.000)	(0.907)
Exp1 ($1 \geq Exp > 0$)	-2.324	-1.502	-1.879	-2.917*	0.112
	(1.821)	(1.505)	(1.716)	(1.703)	(1.714)
Exp2 ($2 \geq Exp > 1$)	-0.572	-2.714**	-3.274**	-4.853***	-2.373*
	(1.557)	(1.202)	(1.592)	(1.193)	(1.244)
Exp5 ($5 \geq Exp > 2$)	-0.433	0.974	-0.866	-1.699	0.318
	(1.145)	(0.993)	(1.156)	(1.042)	(1.023)
Exp10 ($10 \geq Exp > 5$)	-0.093	0.224	-1.074	-2.043**	-0.733
	(0.931)	(0.830)	(0.948)	(0.928)	(0.849)
Constant	-1.901**	-6.422***	-3.222***	-3.820***	-6.707***
	(0.816)	(0.667)	(0.843)	(0.814)	(0.782)
Mean (log-ratio)	-2.381	-6.158	-3.982	-5.687	-6.596
N	327	327	327	327	327

A.11 Documents with the Highest Scores across Categories for Question *classroom management*: "Describe your classroom management style. Why have you found this style to be effective?"

Topic Abbreviations:

- **ASA**: Authoritative Approach for Structure and Accountability
- **SCE**: Student-Centered Engagement for Empowering Relationships
- **PR**: Positive Reinforcement Strategies and Systems
- **MBE**: Modeling Positive Behavior by Example

Table A.22: Documents with the Highest Scores across Categories for Question *classroom management*

ID	Response	ASA	SCE	PR	MBE
894020	Authoritative. Well because it comes closest matching a most students appropriate behavior.	**1**	0	0	0
651896	State my expectation for everyone to follow.	**1**	0	0	0
1110330	The best management style that I am using is that is the authoritative approach style because it is the one most closely associated with appropriate student behaviors.	**1**	0	0	0
6002	I manage my classes with clear expectations and concise procedures.	**0.8**	0	0	0.2
669077	My classroom management style is a classroom that is structure, organized, high expectations, and consistent.	**0.8**	0.2	0	0
17271	Learner Centered approach. This ensures effective learning since the learners are al to build and share knowledge on their own, with me as the facilitator. this approach ensures that learners don't forget their concepts built by themselves.	0	**1**	0	0
19425	I believe in building trust with my students. I spend time talking with them individually when I am able to. I learn what each one likes and dislikes. I take time to listen to their concerns and show respect for every student.	0	**1**	0	0
33216	Letting students lead feel some sort of empowerment is something I've used and is effective. It leads to excitement to be in class and encourages participation.	0	**1**	0	0
37503	Finding how each student learns the best as individuals helps them retain the knowledge being taught even more.	0	**1**	0	0
61086	Classroom management is more about relationships and less about rules.	0	**1**	0	0

Continued on next page

Table A.22: Documents with the Highest Scores across Categories for Question *classroom management* (continued)

ID	Response	ASA	SCE	PR	MBE
21942	Positive Reinforcement. Yes, it is effective because the students like to be appreciated and rewarded for their good behavior.	0	0	**1**	0
596197	Positive reinforcement where students are praised for the great choices they are making. This is effective because all students want that praise and strive to make better choices.	0	0	**1**	0
347538	I like the style of positive reinforcement. This system works because all students can benefit from it. Often times bad behavior gets all the attention. If positive reinforcement is used all students are recognized. It also gives students a chance to change their behavior and get back on track.	0	0	**1**	0
343608	Classroom Dojo. Fake money awards system (willing to discuss).	0	0	**1**	0
497101	I use a star chart, if a student is well behaved everyday of the week, they will receive a star next to their name. At the end of that week the top 3 learners with the most stars receives a certificate for good behavior.	0	0	**1**	0
734966	I like to use the Model Idea Behavior. I feel that in order to earn respect, you have to give respect. I find this works and has shown to be a great help in managing my classroom. I like to show my class first hand how to do techniques in physical education that shows them that I lead by example.	0	0	0	**1**
583049	I believe in leading by example, and teaching is leading with an educational aim. I am the kind of teacher/leader that gets it done with you.	0	0	0	**1**
951498	I have found that leading by example is the best approach however there are certain scenarios when a child will need you to get down to their level.	0	0	0	**1**
960548	I lead from the front I show the kids what to do and help them understand why they are doing what they are doing. To be successful.	0	0	0	**1**
1160443	Modeling. I find it be effective mainly due to response. If I am not showing them how to act, speak, and do, I can not expect them to do the same. Communicating constantly with my students, teachers, parents, and so forth opens any barriers that may be in place.	0	0	0	**1**

Notes: Topic abbreviations: AUTH = Authoritative Approach for Structure and Accountability; SCE = Student-Centered Engagement for Empowering Relationships; PR = Positive Reinforcement Strategies and Systems; MBE = Modeling Positive Behavior by Example. The highest score in each row is bolded.