Enhancing Interview Protocols: Topic Modeling as a Content Validity Technique

Constanza Mardones-Segovia[1]

Tamlyn Lahoud[2]

Cheng Tang[2]

Yaxuan Yang[2]

Shiyu Wang[2]

Allan Cohen[2]

Kun Wang[3]

Chandra Orrill[3]

Rachel Brown[4]

[1]Department of Psychiatry, University of California San Diego.

[2]Department of Educational Psychology, and The University of Georgia.

[3]Rethink Learning Labs.

[4]The Pennsylvania State University.

Affiliation

Author Note

Correspondence concerning this article should be addressed to Constanza Mardones-Segovia, 350 Dickinson St, Suite 3-325, University of California San Diego, San Diego, CA 92103. E-mail: cmardonessegovia@health.ucsd.edu.

The authors have no conflicts of interest to declare.

Abstract

This study proposes a data-driven framework for collecting content and internal structure validity evidence from text-based assessments using Natural Language Processing and Structural Topic Model (STM). A case study of teachers' written responses to an interview protocol illustrates the framework's application. Results show that STM provides a promising approach for generating quantitative evidence from textual assessment by identifying item-construct alignments and uncovering potential sources of construct irrelevant variance. This framework offers practical and data-based inside that support the design and refinement of text-based instruments in educational settings.

Keywords: validity, constructed-response items, natural language model.

Enhancing Interview Protocols: Topic Modeling as a Content Validity Technique

1. Introduction

Constructed-response (CR) items, such as essays, short-response answers, and interview questions, have shown significant utility in quantitative and qualitative research. In quantitative studies, answers to CR items elicit higher-order thinking (Kuechler & Simkin, 2010), facilitate divergent thinking (Guilford, 1957), reduce random guessing (Haladyna & Rodriguez, 2013), and provide a method to evaluate examinees' ability to organize ideas coherently (Kuechler & Simkin, 2010). In qualitative studies, responses in interview protocols help to explore and understand interviewers' beliefs and perspectives in-depth towards a certain phenomenon, enabling rich interpretations of how interviewers assign meaning to their social experiences (Dunwoodie, Macaulay, & Newman, 2023).

An important aspect when measuring latent variables (e.g., abilities, beliefs, etc.) is to ensure that the items or questions reflect the intended construct and that the interpretations obtained from tests, assessments, or interviews (hereafter called instruments) are valid for a specific purpose (Zumbo, 2006). Thus, the information derived from items depends on the quality of the instrument and the validity evidence supporting its use (S. Sireci & Benítez, 2023). For instance, a key question is whether items on an IQ test genuinely reflect a person's intelligence. As such, validity must be carefully considered by both assessment developers and stakeholders who rely on these assessments within their specific contexts (Phakiti & Isaacs, 2021).

In general, we provide evidence of content validity using ratings from expert judges and construct validity using methods such as factor analysis, structural equation model (SEM), and item response theory (IRT), which rely on numerical scores. However, CR items contain rich information that can be lost when transformed into numerical representation. Similarly, interview data is typically analyzed with qualitative techniques such as content analysis, which uncovers patterns that help understand a social group. Yet,

due to its unstructured nature, interview data may introduce unanticipated content, making it difficulty to analyze with traditional quantitative methods.

New advances in artificial intelligence have provided methods for analyzing textual data quantitatively while maintaining the meaning of words. For example, topic models have been applied to analyze students' thinking and reasoning in answers to CR items (for each item). However, no information is available yet about how to apply traditional validity evidence techniques when using NLP methods.

This article proposes using STM as a data-driven approach to provide validity evidence of test content and internal structure, offering a more automatic and interpretable method for assessing validity in text-based assessments.

2. Validity Evidence with Traditional Methods

Assessing the degree to which each item in a test reflects the intended construct is a critical aspect of measurement, as the information derived from test items depends on the quality of the test and the validity evidence supporting its use (S. Sireci & Benítez, 2023). Validity is a unitary concept that reflects the extent to which the evidence and theory support the test's score interpretations (S. G. Sireci, 1998). It is not an inherent property of a test but rather an ongoing process of collecting evidence.

The Standards from American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) outline five sources of validity: (1) test content, (2) response process, (3) internal structure, (4) relationship to other variables, and (5) testing consequences (American Educational Research Association & National Council on Measurement in Education, 2014). Among these, validity evidence based on test content, also called content validity, is a critical measurement aspect. It evaluates the degree to which each item in a test reflects the intended construct (S. G. Sireci, 1998).

Traditionally, content validity evidence is provided by expert judges who evaluate how well each item is *representative* and *comprehensive* to measure the intended construct.

In other words, the degree to which each item is relevant to the construct (Almeida & Xexéo, 2019) and the degree to which the test covers the full scope of the intended construct (S. Sireci & Benítez, 2023). To this end, the panel of experts provides both quantitative and qualitative evaluations, considering whether the items align with the study purpose, measure the intended content, and are clearly formulated. Their evaluations are then analyzed using quantitative techniques to ensure agreement among the experts (Spoto, Nucci, Prunetti, & Vicovaro, 2023).

In qualitative research, particularly when using interviews, validity is conceptualized differently. While test items in quantitative assessments are evaluated for their *relevance* and *comprehensiveness*, interview questions are assessed based on their ability to elicit meaningful and in-depth responses that align with the study purpose. Applying the same procedure from quantitative research to obtain content validity evidence presents several challenges. Interview questions are often more flexible, allowing variations in wording or follow-up questions to clarify participants' responses. Moreover, the open-ended nature of interview responses can lead to the measurement of unanticipated content that traditional quantitative methods might fail to capture. This flexibility and unstructured response data make traditional quantitative methods less effective in evaluating content validity in this setting (Torlig, Junior, Fujihara, Demo, & Montezano, 2022).

3. Topic Models

Before explaining how STM can be used as a data-driven method to obtain validity evidence, the following section provides an overview of topic models, starting with the simplest model, Latent Dirichlet Allocation, and proceeding to the development of STM.

Topic models are machine learning techniques in Natural Language Processing (NLP) designed to identify, categorize, and summarize large collections of textual documents into a latent topic structure (Blei, 2012). In general, topic models have been applied to individual items to retrieve additional information from CR items or, in the case of interview data, to summarize interviewees' beliefs in each question. For instance,

researchers have applied topic models to analyze CR items in educational assessments, finding that the topics identified by the model reflected key ideas emphasized in the expert scoring guidelines (Choi et al., 2019). Similarly, they have been used to examine open-ended responses about the education system, with the resulting topics capturing commonly held concerns among students, teachers, and parents (Cifuentes & Olarte, 2023). In higher education, topic models have also been used to summarize themes in student course evaluations, demonstrating their utility for large-scale interpretation of feedback (Sun & Yan, 2023).

Unlike traditional quantitative methods, which assume examinees' scores on test items (categorical, ordinal, or continuous) as the observed variables, or qualitative methods, which treat interviewees' textual responses as the observed units, topic models assume words within a document as the observed variables. Specifically, topic models assume a corpus M is a collection of D documents. Each document d consists of a sequence of N_d words, where each word $w_{d,n}$ in a document d at position n is an element of a finite vocabulary of size V ($w_{d,n} \in \{1, 2, \ldots, V\}$).

Some of the most widely used topic models in education and social sciences are the Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) and STM (Roberts et al., 2014). Both are unsupervised topic models, meaning the latent topic structure is unknown beforehand, similar in nature to an exploratory factor analysis (EFA). For example, both evaluate multiple candidate models with different numbers of topics or factors, and the best-fitted model is selected for interpretation. However, although topic models share similarities with EFA, they also have distinct differences. For instance, topic models assume that both observed and latent variables are categorical and belong to a family of probabilistic models known as mixed membership models. This implies that documents can be associated with multiple topics, such that a document may contain words related to more than one topic. Additionally, topic models assume that examinees' answers generate latent topics, and these influence the observed words. In contrast, EFA assumes that the

item responses are indicators of the underlying factor structure.

The simplest unsupervised topic model is LDA, which estimates the number and content of topics based only on the words associated with each document. On the contrary, STM generalizes LDA by including additional information to estimate their effect in the prevalence and content of topics (Abraham, Mardones-Segovia, Sarles-Whittlesey, & Cohen, 2024). Each of them is detailed below.

2.1 Latent Dirichlet Allocation

LDA is a generative probabilistic model that estimates the probability distribution over words (observed variables) and topics (latent variables). It models how each document d is generated as a mixture of topics, with each topic k represented as a distribution over words (Blei et al., 2003).

More formally, LDA is defined by three main parameters. Each topic $k \in \{1, ..., K\}$ is represented by a word-topic distribution denoted as $\vec{\phi}_k = [\phi_{k,1}, ..., \phi_{k,V}]$, which indicates the probability of words occurring in each k topic. These distributions form a $K \times V$ matrix, where each $\phi_{k,v}$ entry denotes the probability of the vth word belonging to the kth topic. Each document d is represented by a document-topic distribution (also referred to as topic proportion), denoted as $\vec{\theta}_d = [\theta_{d,1}, ..., \theta_{d,K}]^T$, which represents the probability distribution over topics for document d. These distributions form a $D \times K$ matrix, where each entry $\theta_{d,k}$ indicates the probability that the dth document is associated with the content of the kth topic. Since ϕ_k and θ_d are probability distributions, they satisfy the following constrained: $\sum_{v=1}^{V} \phi_{k,v} = 1$, $\forall k \in \{1, \ldots, K\}$ and $\sum_{k=1}^{K} \theta_{d,k} = 1$, $\forall d \in \{1, \ldots, D\}$. Finally, the topic assignments, denoted as $z_{d,n}$, indicate the estimated topic membership of each word n within a document d.

2.2 Structural Topic Model

The Structural Topic Model (Roberts et al., 2014) generalizes LDA by incorporating document-level information, such as author characteristics, time, or other contextual information, as covariates in the topic modeling process. Rather than assuming fixed

distributions of topics within documents and words within topics, STM allows these distributions to vary based on these external variables. This flexibility enables researchers to investigate how covariates influence the prevalence of topics across documents (topical prevalence) and the language used within topics (topical content), offering detailed information regarding the relationship between textual responses and contextual variables.

Formally, STM retains the main LDA parameters, where $\vec{\theta}$ is a vector representing the distribution of topics in each document d, and $\vec{\phi}$ represents the word-topic distributions. However, instead of assuming that both are drawn from a Dirichlet distribution with hyperparameters α and β, respectively, STM models them as functions of the covariates (Roberts, Stewart, & Tingley, 2019). Specifically, $\vec{\theta}$ is modeled through a logistic regression, where β denotes the regression coefficient capturing the effect of covariates on topic proportions, and Σ represents the covariance matrix showing the relationship among topics. In contrast, $\vec{\phi}$ is modeled using an additive approach, where the word distributions are adjusted based on the external variable to account for variations in topical content.

The generative process in STM assumes that, for each word w in a document d, a topic assignment z is drawn from the document-topic distribution $\vec{\theta}$, such that $z \sim Multinomial(\vec{\theta})$. On the contrary, the word-topic distribution is modeled as $\vec{\phi} \propto \exp(m + \kappa)$, with m being the baseline word frequency and κ reflecting the topic-specific variations. These modeling variations enable (1) topics to correlate, (2) each examinee's answer to have a unique distribution over topics, and (3) the content to vary based on covariates (Roberts et al., 2014).

As previously stated, topic models have been used in formative assessments with mixed-format or CR tests, surveys containing CR items, and interview data to analyze examinees' written responses at the document level (e.g., Cardozo-Gaibisso, Kim, Buxton, & Cohen, 2019). This allows researchers to examine the examinees' reasoning and thought processes for each item, with and without considering the influence of external variables.

The STM versatility allows us to extend its applicability beyond individual items or

documents, enabling a test-level approach where examinees' responses to all items are analyzed jointly. By incorporating item identifications as covariates, we can evaluate how individual items influence the distribution of topics across responses. This approach provides a novel framework for exploring content and construct validity for written response items, as described in the following section.

4. STM as a Data-Driven Validity Framework

This section introduces a novel methodological framework to provide validity evidence regarding test content and internal structure for assessments composed of written responses to CR items or interview data. Each approach will be explained below.

Compared to traditional strategies to gather validity evidence, which assume that the constructs measured by a test are predetermined, this data-driven approach can reveal unanticipated content that conventional quantitative methods might fail to capture, identifying irrelevant or unintended constructs that could introduce bias.

By incorporating item identifications as covariates for the topic proportions, STM allows us to achieve four key objectives. First, it identifies latent topics that summarize examinees' written answers to test questions. This allows for exploring latent topics without assuming the underlying constructs beforehand. Second, it estimates how well a test is both *relevant* and *comprehensive* for its intended purpose. The topic prevalence results indicate the degree to which each item elicits responses aligned with each topic (Almanasreh, Moles, & Chen, 2019), reflecting whether the item responses are pertinent to the measured construct. Third, it facilitates evaluating whether questions comprehensively cover all intended topics or if specific areas are underrepresented, indicating content gaps. Finally, this approach can also detect instances where responses reflect unintended constructs, indicating potential construct-irrelevant variance. Together, STM offers a comprehensive framework for gathering content validity evidence.

Beyond evaluating content validity, STM's regression framework also facilitates preliminary assessment of construct validity. Specifically, the regression coefficients

associated with each item reflect the strength between individual items and the latent topic. In a Bayesian framework, the posterior mean of the regression coefficients estimates the relationship between item and topic, while its corresponding credible interval measures the uncertainty around this estimate. These results will provide insight into whether items adequately represent the intended construct. Conceptually, the posterior mean is similar to factor loadings in an exploratory factor analysis, which measure the extent to which an item aligns with an underlying construct.

Two aspects influence how the content and latent structure of responses are interpreted: (1) *item complexity* and (2) *construct relationship*. Items can be classified as either simple or complex depending on the number of latent variables they are intended to measure. Simple items are designed to elicit responses aligned with a single content, whereas complex items can simultaneously elicit answers from more than one content. Likewise, constructs may be independent, correlated, or hierarchically related, depending on the nature of their conceptual relationships. Our framework allows for exploring item complexity through the STM's mixed-membership structure and examining construction relationship through a posterior topic correlation analysis. However, STM does not directly support exploring hierarchical contents in its current form. See section 4.1.6. for detailed information.

The STM as a data-driven approach includes five main tasks: (1) Data preprocessing, (2) STM configuration, (3) selecting the number of topics, (4) interpreting and labeling topics, and (6) providing data-driven validity evidence. Each of them is explained below.

4.1.1. Data Preprocessing

STM takes as input a Document-Term Matrix denoted as $\boldsymbol{DTM}$, with dimensions $D \times V$. Each DTM_{dv} entry indicates the frequency of word v in document d. Within the NLP framework, this corresponds to the bag-of-word feature extraction method, where the sequence or order of words is disregarded, and only word frequency is considered in the

topic modeling process (Jurafsky & Martin, 2023).

Different NLP techniques are applied to pre-process examinees' written responses and convert them into a $\boldsymbol{DTM}$. This process includes converting words to lowercase, correcting misspellings mistakes, removing punctuations and high frequent words that do not carry relevant information to interpret the results (e.g., 'the', 'a', 'of'), normalizing the words to a common root (e.g., the words 'jump' and 'jumping' portray the same meaning but in difference tenses), and tokanizing or separating words into meaningful units. After tokenizing words, these are converted as a $\boldsymbol{DTM}$.

4.1.2. STM Configuration

The proposed method uses item IDs to evaluate their effect on topic prevalence. To account for the fact that an examinee may respond to multiple items, examinee ID is also included as an additional covariate to control for non-independence. In this approach, covariates are specified only for the topic proportion, meaning that the word-topic distribution ϕ_k is estimated directly from the observed words' frequencies, similar to the standard approach used in LDA.

Let $j \in \{1, \ldots, J\}$ denote examinees' ID and $i \in \{1, \ldots I\}$ represent the item's ID. Each document d corresponds to one written response from examinee j to item i. In the STM framework, these categorical variables are internally converted into binary indicator variables, omitting one reference category to avoid multicollinearity. Accordingly, STM defines a design matrix $\boldsymbol{X}$ of dimensions $D \times P$, where $P = (I + J) - 2$ represents the number of dummy-coded covariates (Roberts et al., 2014).

Given the selected topic structure, the topic proportion vector for each document d is modeled as: $\theta_d \sim LogisticNormal(\boldsymbol{X}_d\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{X}_d$ is a row vector containing the dummy-coded covariates, $\boldsymbol{\Sigma}$ is the covariance matrix of topics, and $\boldsymbol{\beta}$ is the regression coefficients matrix of dimension $P \times K$, such that each entry β_{pk} reflects the effect of examinee j and item i on the prevalence of topic k. These effects are in a logit scale, meaning each β_{pk} represents the changes in the log-odds of topic k being discussed in a

document when holding the proportion of the other topics constant.

In this model configuration, the mean vector of the document-topic distribution is conditioned on both examinee and item IDs. However, as previously stated, examinee IDs are included only as a control variable to account for the non-independence of responses from the same examinee. Thus, only the item coefficients are used for validity evidence.

STM includes two main strategies to start the model: the LDA and the Spectral initialization. The LDA initialization method uses the collapsed Gibbs sampling for Latent Dirichlet Allocation (LDA) to produce initial topic-word distributions before proceeding with STM's variational inference algorithm (Blei et al., 2003). On the contrary, the Spectral initialization method leverages non-negative matrix factorization of the word co-occurrence matrix to provide globally consistent initial parameter values, resulting in more stable and reproducible topic solutions Roberts et al. (2019).

To our knowledge, no prior study has evaluated the optimal STM configurations in assessment data. Thereby, this study evaluates the initialization strategies when presenting a case study in section 5.

4.1.3. Model Selection

After converting examinees' written responses to a numerical format, we specify the topic structure and select the best-fitted model. Typically, studies in education run between 2 and 10 candidate topic models (e.g., Cardozo-Gaibisso et al., 2019) and evaluate them based on two coherence metrics. Semantic coherence measures the tendency of high-probability words within a topic to co-occur together, indicating internal consistency of topics, while exclusivity assesses the degree to which high-probability words in a topic are exclusive to that topic rather than appearing with high probability across multiple topics. The best latent topic structure, then, is defined as the model that maximizes both, semantic coherence and exclusivity (Abraham et al., 2024).

Semantic coherence measures the tendency of high-probability words within a topic to co-occur together, indicating internal consistency of topics, while exclusivity assesses the

degree to which high-probability words in a topic are exclusive to that topic rather than appearing with high probability across multiple topics. The optimal model should balance these two metrics to reach for an optimized results.

4.1.4. Interpreting and Labeling Topics

After selecting the best-fitted model, STM estimates the document-topic and word-topic distributions. The most probable documents and words within a topic form the basis for interpreting and labeling the latent topic structure.

Specifically, ChatGPT-4o interprets the content of each topic using the most representative documents (examinees' answers) for each topic. These interpretations are then reviewed by expert judges to evaluate the accuracy of the interpretations. For instance, consider examinees one, two, and three, whose topic proportions are: $\vec{\theta}_1 = [0.94, 0.03, 0.03]$, $\vec{\theta}_2 = [0.02, 0.95, 0.03]$, and $\vec{\theta}_3 = [0.01, 0.02, 0.97]$, respectively. This information indicates that 94% of the content of examinee one corresponds to topic 1, 95% of the content of examinee two corresponds to topic 2, and 97% of the content of examinee three corresponds to topic 3. Accordingly, ChatGPT uses the responses of examinees one, two, and three to interpret and label topics 1, 2, and 3, respectively. By selecting highly probable documents for each topic, ChatGPT examines the common words across responses and provides an initial interpretation to ease the expert judges' task.

4.1.5. Data-driven Validity Evidence

After fitting the regression model for the selected topic model, two main outputs are obtained: (1) the posterior mean topic proportion for each item in a given topic and (2) its 95% credible intervals. The first output is used to gather content validity evidence, while the second supports initial internal structure validity evidence by reflecting the uncertainty around an item-topic association. Each of them is explained below.

4.1.5.1. Evidence of content validity

For simplicity, the posterior mean topic proportions for each topic across items can be visualized using bar plots, where the $x-$axis represents the item ID and the $y-$axis

shows the posterior mean topic proportions.

The first piece of evidence comes from the number and interpretation of topics extracted. These topics may align with, differ from, or extend the constructs the instrument was intended to measure. In theory, a construct's conceptual definition should closely resemble its corresponding topic's definition (American Educational Research Association & National Council on Measurement in Education, 2014). For example, suppose the instrument was designed to measure proportional reasoning, fractional reasoning, and problem solving, but the answers to the test were clustered in only two topics. In that case, this may indicate that the items did not fully cover the intended construct. On the contrary, if the responses reflected four topics, it could indicate that the items elicit unintended constructs, showing a potential source of measurement bias. Expert judges can be used to evaluate the interpretation of these topics and assess their alignment with the conceptual definition of the intended construct.

The second piece of evidence comes from identifying the likelihood of each item in each topic, as estimated by the posterior topic proportion. As noted before, the interpretation of the item-topic relationship depends on the combination of item complexity and construct relationship, with four cases:

1. *Independent constructs with simple items.* When an item is designed to measure a specific content, the estimated topic proportion for its corresponding topic should be close to one, and therefore, zero for others. For example, if a simple item elicits response patterns from more than one topic, it may indicate that the item prompt was ambiguous or that the item measures unintended content, thereby raising concerns regarding its quality. Conversely, if the item elicits responses primarily aligned with a single topic, it provides evidence of its clarity and content alignment.

2. *Independent constructs with complex items.* For items designed to measure multiple aspects of a construct, the estimated topic proportion is expected to be more evenly distributed across the relevant topics. Accordingly, if an item produces answers

related to a mixture of topics with one being more predominant, it may reveal an imbalance in how those different concepts are being measured.

3. *Related constructs with simple items.* When an item is designed to measure a single construct that is conceptually related to another, the topic proportion may show some overlap, such that the answers to the item predominantly include words related to one topic. However, if the proportions are more evenly distributed across topics, it may indicate that the item is measuring additional constructs.

4. *Related constructs with complex items.* When an item is designed to measure two or more related constructs, the estimated topic proportion should be balanced across the topics. However, if the topic proportions are heavily skewed toward one topic, it may suggest that the item is not capturing the intended constructs equally.

The final piece of evidence comes from the representativeness of items across topics. In traditional test theory, a concept, skill, or ability requires a sufficient number of items to measure a given concept accurately. Likewise, each concept should be measured using a similar number of items of similar quality (S. G. Sireci, 1998). For example, Diagnostic Classification Models suggest that a test must include between three and five items for each attribute and a balanced Q-matrix to provide stable parameter estimates (Bradshaw & Madison, 2016). In the STM framework, we can observe the number of items that load higher for each topic and assess their representativeness and balance by comparing the number of items across topics. Likewise, we can observe the topic density by counting the most predominant documents for each topic. If a topic is less discussed across documents, it can indicate that the items did not fully capture the intended topic.

4.1.5.2. Evidence of internal structure

Besides using the posterior mean, i.e., the estimated mean topic proportion, a 95% credible interval can be utilized as a measure of uncertainty around this estimate, offering evidence of construct validity. In factor analysis research, loadings ≥ 0.7 typically represent

a strong association between an item and a factor (e.g., Tabachnick & Fidell, 2007). Because of their similarities, this study uses the same cutoff criterion to interpret the item-topic alignment.

Evidence of construct validity will vary depending on the four cases detailed before:

1. *Independent constructs with simple items.* When an item is designed to measure a single construct that is unrelated to others, we expect the posterior topic proportion to be high (≥ 0.70) for one topic and close to zero for all others. A narrow, credible interval (e.g., a range of 0.10 or less) around the dominant topic provides stronger evidence that the item consistently elicits responses aligned with the intended construct. If the item shows moderate to high proportions across multiple unrelated topics, or if the intervals overlap, this may suggest item ambiguity, unintended construct activation, or poor alignment. Conversely, if an item has a low posterior mean and narrow interval for a topic it was not designed to measure, this supports construct validity by showing discriminant evidence.

2. *Independent constructs with complex items.* When an item is designed to measure two or more distinct (uncorrelated) constructs, we expect posterior topic proportions to be moderately distributed across the corresponding topics, each with narrow and non-overlapping credible intervals. Non-overlapping intervals suggest that the item certainty elicits words related to each construct. If one topic disproportionately dominates (e.g., ≥ 0.70), or if the credible intervals overlap substantially, this may indicate that one construct is underrepresented, the item is imbalanced, or there is uncertainty in how constructs are elicited.

3. *Related constructs with simple items.* When an item is designed to measure a single construct that is conceptually related to others, the posterior topic proportion should be high for the intended construct. Compared to the other cases, credible intervals could potentially overlap. However, if topic proportions are more evenly distributed

across topics or the credible intervals for different topics overlap substantially, this may suggest that the item unintentionally elicits responses aligned with multiple constructs, weakening construct clarity.

4. *Related constructs with complex items.* When an item is complex and the constructs are related, we expect a more evenly distributed topic proportion with overlapping credible intervals. In this case, the overlapping reflects the integration of concepts. Conversely, if the answers to the item predominantly elicit words related to one topic or the intervals do not overlap, it may suggest that the item is not capturing the shared construct.

As noted before, because our proposed model is conceptually similar to factor analysis, we can visualize the topic structure by creating a diagram that shows which items elicit answers related to a topic. In our proposed framework, however, item responses generate the documents, each of which is a mixture of topics, and these topics are a mixture of the observed words (item responses $\rightarrow$ documents $\rightarrow$ topics $\rightarrow$ words).

We can visualize the models' results as a directed acyclic graph (DAG), effectively showing the relationship between items and topics. In this diagram, items are observed nodes that elicit the responses related to a topic and topics are latent nodes representing the concepts or themes that summarize a collection of examinees' written responses. The edges, represented by arrows, connect the items to the topic. An item-topic edge denotes the degree of alignment between an item and topic and is quantified by the posterior topic proportion. The corresponding credible interval, shown in square brackets, reflect the uncertainty around the estimate.

5. Case Study

This section includes a case study to illustrate step-by-step the methodological framework. Because no prior study has evaluated the optimal STM configurations in assessment data and the number of topics is unknown, this study evaluates STM's results

by manipulating two factors: (1) initialization strategy and (2) number of candidate models. Each of them is detailed in sections 5.3 and 5.4.

5.1 Case Background

The proposed framework was applied to a dataset including responses from 16 middle-grade math teachers from New Jersey and Florida. These responses were collected via a semi-structured interview composed of 11 main questions (or prompts) including 36 sub-questions in total. This interview sought to capture their thinking processes when solving fraction problems before participating in a professional development (PD) intervention aimed at expanding their knowledge of fractions and proportional reasoning. Due to time constraints, each teacher responded to approximately seven questions, thereby resulting in a corpus of 502 documents, where each document corresponded to a teacher's response to specific questions and sub-questions.

The interview protocol aimed to measure four mathematical concepts: (1) Referent Unit, (2) Invariance, (3) Covariance, and (4) Quantity. Reference Unit refers to the whole quantity to a fraction or ratio. Invariance is the property that remains unchanged under certain operations. Covariation shows how two quantities vary with each other. Quantity is the measurement of objects or phenomena.

Table 1 shows sub-items distribution across themes. A value of 1 indicates that the sub-question was designed to measure a theme, while zero indicates otherwise. The column labeled Theme Total represents the number of themes measured by each item. Sub-items with values greater than 1 in this column are considered complex, whereas those with a value of 1 are considered simple. The row labeled Item Total shows the total number of sub-questions associated with each theme. Overall, the distribution of sub-questions across themes is uneven, with Referent Unit being measured by 28 sub-questions and Covariance by only two sub-questions. Additionally, 16 of the 36 sub-questions measured more than one theme, indicating a substantial proportion of complex items.

Table 1

Sub-items by Themes

Sub-items	Covariance	Invariance	Quantity	Referent Unit	Themes Total
1.1	0	0	0	1	1
1.2	0	1	0	1	2
1.3	0	0	0	1	1
1.4	0	0	1	1	2
1.5	0	0	1	1	2
1.6	0	0	0	1	1
2.1	0	1	0	1	2
2.2	0	1	0	0	1
3.1	0	1	1	0	2
3.2	0	1	0	0	1
4.1	0	1	1	1	3
4.2	0	1	1	1	3
4.3	0	1	1	1	3
4.4	0	0	0	1	1
5.1	0	0	1	1	2
5.2	0	1	0	1	2
5.3	0	1	0	1	2
5.4	0	1	0	1	2
6.1	0	0	0	1	1
6.2	0	0	0	1	1
6.3	0	0	0	1	1
6.4	0	0	0	1	1
6.5	0	0	0	1	1
7.1	0	0	1	0	1
7.2	0	0	1	0	1
7.3	0	0	1	0	1
8.1	1	1	0	1	3
8.2	0	0	0	1	1
8.3	1	1	0	0	2
9.1	0	1	0	1	2
9.2	0	0	1	1	2
10.1	0	0	0	1	1
10.2	0	1	0	0	1
10.3	0	0	0	1	1
11.1	0	0	0	1	1
11.2	0	0	0	1	1
Items Total	2	15	11	28	56

5.2. Data Preprocessing

Before applying the STM, the raw text data needs to be systematically preprocessed to prepare it for subsequent analysis. The interview data was automatically transcribed from the audio files, which could potentially lead to the situations of extraneous whitespaces, typos, and extra punctuation. In order to standardize the text, the data went through a series of normalization procedures. Punctuations, numbers, and extraneous whitespace were removed. Misspellings are replaced. Moreover, all texts are converted to lowercase to avoid inconsistencies arising from case differences.

In this study, the stopword list includes both standard English stopwords and a custom list specifically created based on the dataset. Additional stopwords were classified into 11 categories according to their meanings: agreement, uncertainty, exclamations, casual expressions, time and condition, gratitude, quantity, miscellaneous, action, states, and others. These filler words and irrelevant phrases can introduce noise into the analysis. Therefore, a comprehensive stopword list was developed based on these criteria, and all identified stopwords were removed accordingly.

Moreover, some phrases were modified in the preprocessing step. The topic modeling can identify phrases with more than one word as multiple words, which could cause confusion. For example, during the interview, a few teachers used phrases like 'not equivalent' or 'not equal'. These phrases will be treated as two individual words by topic modeling, which could cause confusion. To avoid the separation, such phrases are transferred into a single word by removing the whitespace between the words. In this case, 'not equal' is recorded as 'notequal'.

Next, lemmatization was applied to reduce inflectional and derivational forms of words to their base or dictionary form. This process allows us to preserve the semantic meaning of words while reducing lexical variation. During the process, all text is converted to lowercase to ensure consistency. Then various word forms are transformed to their base form while preserving special mathematical terminology relevant to fractions and rational

numbers. Following lemmatization, we implemented a second stage of preprocessing to remove stopwords and ensure that the lemmatization process had not introduced unintended words.

Before the data was cleaned, the corpus contained 4,488 unique words and 502 documents with a total number. However, due to the suboptimal quality in portions of the interviews, the qualitative team decided to exclude sub-questions 1.1, 1.2, 1.6, and all those under sections 3, 5, and 10. Thus these items were not included in the further coding process. Following the exclusion of these items and the preprocessing step, the corpus reduced to 1,129 unique words and 321 documents. See Table 2 for more details.

Table 2
Comparison of Descriptive Statistics Before and After Preprocessing

State	Vocabulary of Unique Words	Documents	Total Words	Average Document Length	Standard Deviation
Before Preprocessing	4488	502	59298	118.13	88.83
After Preprocessing	1129	321	16512	51.44	37.51

5.3 STM configuration

STM included interview sub-questions and participant IDs as document-level covariates, enabling topic prevalence variations based on the specific question being discussed.

As mentioned, STM includes two main strategies to start the model: the LDA and the Spectral initialization. We evaluated four strategies: (1) spectral initialization and LDA initialization with (2) $\alpha = 0.5$ and $\beta = 0.05$, (3) $\alpha = 1$ and $\beta = 1$ Mardones-Segovia, Choi, Hong, Wheeler, and Cohen (2022), and (4) $\alpha = 50/\mathrm{T}$ and $\beta = 0.01$, where T represents the number of topics Steyvers and Griffiths (2007).For the models with LDA initialization, we also set the burn-in period as 10,000 iterations and the number of iterations post-burn-in as 15,000 iterations.

5.4. Model selection

For each initialization strategy, we estimated a set of candidate topic models ranging from 2 to 10 topics as suggested in previous studies (e.g., Cardozo-Gaibisso et al., 2019). Thereby, we studied 32 possible parameters' combinations in total.

The best-fitting model was selected using semantic coherence and exclusivity. Figure 1 shows a scatterplot with the model selection results for the 32 tested models. These results suggest that a 3-topic, 5-topic, 6-topic, and 10-topic model with spectral initialization, and a 6-topic model with LDA initialization (α=1, β=1) stand out for their balance in both semantic coherence and exclusivity.

To further investigate which model performs best, we applied Min-Max Normalization to measure the two metrics in the same scale (Patro & Sahu, 2015). After normalization, equal weights (0.5) were assigned to each metric to compute a combined score for each model. This step ensures that semantic coherence and exclusivity contribute equally to model evaluation (Singh & Singh, 2022).

The results shown in Table 3 suggest that models with spectral initialization have better performance than those using LDA. Among the spectral initialization models, the 10-topic model achieved the highest combined score (0.713), closely followed by the 6-topic (0.705) and 3-topic (0.704) models, with minimal differences among them. Based on these findings, we selected the 10-topic, 6-topic, and 3-topic models from the spectral initialization for further comparison.

Table 3
Normalized Score for Semantic Coherence and Exclusivity

candidate_model	SC	EXC	Score
Spectral-3-topic	1.000	0.409	0.704
Spectral-5-topic	0.592	0.746	0.669
Spectral-6-topic	0.573	0.837	0.705
Spectral-10-topic	0.426	1.000	0.713
LDA-6-topic (α=1, β=1)	0.394	0.818	0.606

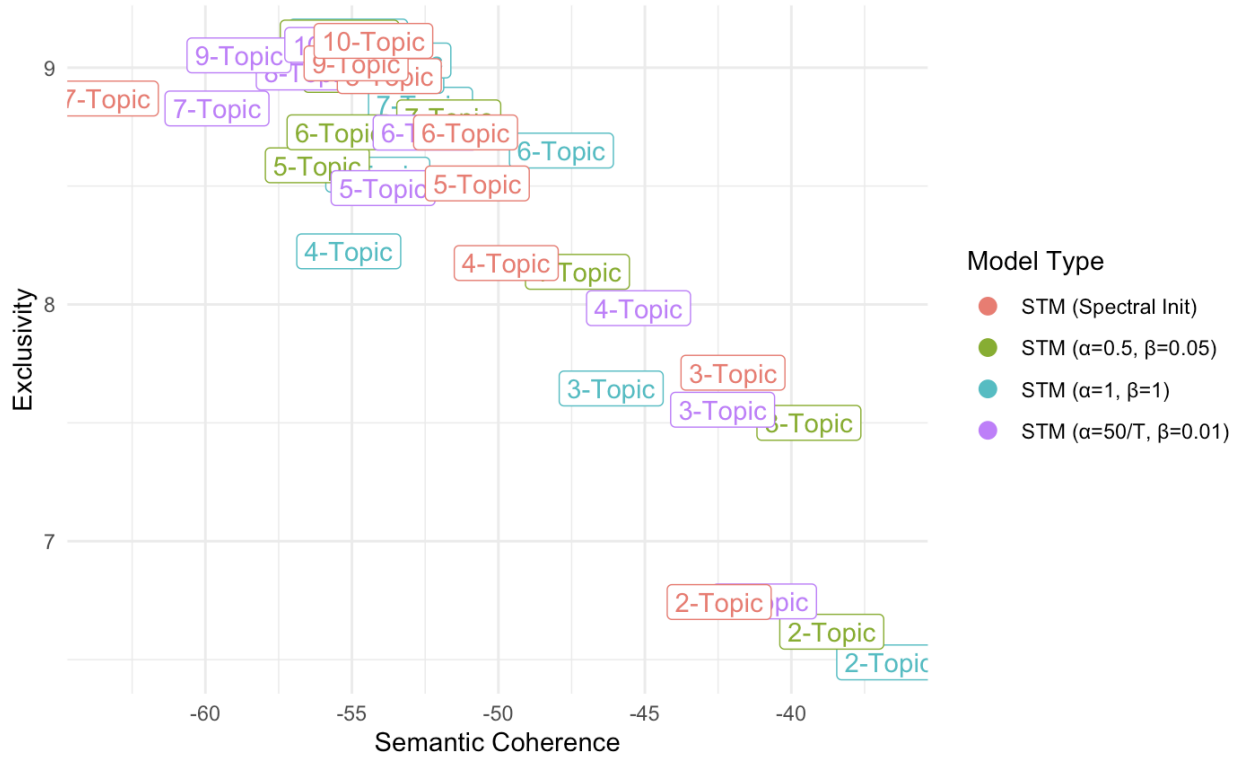

Figure 1
Model Selection

5.5. Interpreting and Labeling topics

For the remaining three models, we use ChatGPT 4o to interpret the top 10 documents most strongly association with each topic. The interpretation results are presented in Tables 4, 5, and 6, respectively along with the corresponding number of documents per dominant topic shown in Figures 2, 3, and 4.

Overall, the 6-topic model offered the best balance between interpretability and granularity, effectively capturing key aspects of teachers' understanding, including number line visualization for fraction division, equivalent fractions and proportional relationships, unit relationships in context, visual analysis of fraction equivalence, area models for multiplication, and contextual fraction problems. This topic structure aligned closely with the original constructs while avoiding the redundancy observed in the 10-topic model, where several topics overlap in their focus on referent units.

Table 4

Large Language Model Interpretation Summary (3-Topic)

Topic	Label	Key Content
1	Contextual Fraction Problem-Solving	Teachers apply the fraction concepts in a contextualized garden scenario. Teachers demonstrate their understanding by determining what fraction of a whole (garden or acre) different sections represent.
2	Analysis of Representational Models for Equivalence and Proportion	Teachers analyze different representations of equivalent fractions and proportional relationships. Teachers evaluate the pedagogical effectiveness of representations, distinguishing between those that better illustrate equivalence versus those that demonstrate proportionality.
3	Analyzing Area Models for Fraction Multiplication	Teachers interpret and critique area models for multiplying fractions. Teachers analyze student work samples, identifying how fractions are represented within area models and evaluating the correctness and clarity of different approaches.

Table 5

Large Language Model Interpretation Summary (6-Topic)

Topic	Label	Key Content
1	Number Line Visualization for Fraction Division	Teachers demonstrate how number lines can be used to visualize fraction division, particularly the relationship between 8/16 and 1/4. They use visual models to help bridge abstract mathematical concepts with concrete representations. Teachers value these visualizations for helping students see relationships between fractions.
2	Equivalent Fractions and Proportional Relationships	Teachers analyze different visual models representing the same mathematical relationship (2/3 = 8/12), distinguishing which representations better demonstrate fraction equivalence versus proportional relationships. This reveals pedagogical content knowledge as they evaluate representations based on their instructional affordances, recognizing that different models serve different pedagogical purposes depending on the concept being taught.
3	Unit Relationships in Garden Context	Teachers reason about fractional relationships between different-sized garden beds, demonstrating understanding of how establishing a referent unit is crucial for determining fractional relationships. They show flexibility by defining different referent units and calculating corresponding relationships, converting between different referent units, and demonstrating grasp of the relative nature of fractions in contextual problems.
4	Identifying Equivalent Fractions in Visual Representations	Teachers analyze diverse visual representations to determine fraction equivalence, demonstrating ability to recognize that different visual models can represent the same fraction value despite varying appearances. They identify which representations are equivalent to 1/2 and which are not, explaining their reasoning process and showing capacity for flexible visual reasoning with fractions across various representations.
5	Area Models for Fraction Multiplication	Teachers engage with area models representing fraction multiplication ($3/4 \times 2/3$), demonstrating varying levels of comfort and familiarity with interpreting these visual representations. Their responses reveal analytical processes as they make sense of how visual models correspond to mathematical operations.
6	Basketball and Pizza Fraction Addition Contexts	Teachers analyze two different situations that appear to involve fraction addition but represent different mathematical concepts: the basketball free throw context ($2/3 + 3/4 = 5/7$) involving combining ratios, and the pizza context ($2/3 + 3/4 = 17/12$) involving standard fraction addition. They identify fundamental differences between these contexts, recognizing distinctions between ratio reasoning and part-whole interpretations.

Table 6

Large Language Model Interpretation Summary (10-Topic)

Topic	Label	Key Content
1	Defining and Applying Referent Units	Teachers engage with using one object as a referent unit to describe the relative sizes of other objects, demonstrating the ability to use a given object (marigolds) as a unit of measurement.
2	Representations of Fraction Multiplication	Teachers evaluate different representations of fraction multiplication ($2/3 \times 6/7$), demonstrating varying levels of familiarity and comfort with different visual models.
3	Number Lines for Fraction Relationships	Teachers interpret how number lines can represent relationships between fractions. The teachers primarily focus on using the visual representation to understand fraction division or to compare equivalent fractions.
4	Analyzing Geometric Representations of Equivalent Fractions	This topic reveals teachers' ability to analyze geometric shapes to identify patterns and relationships in equivalent fractions. Teachers demonstrate skill in recognizing when different visual representations.
5	Units and Wholes in Fraction Contexts	This topic emphasizes teachers' understanding of how the definition of the "whole" or "unit" affects fraction interpretation.
6	Comparing Models for Proportional Relationships	This topic focuses on teachers' comparative evaluation of different models for teaching equivalent fractions versus proportional relationships. Teachers demonstrate knowledge by distinguishing which representations better support different concepts.
7	Analyzing Area Models for Fraction Multiplication	This topic shows teachers critically analyzing an area model for fraction multiplication ($3/4 \times 2/3$). Teachers demonstrate their ability to evaluate representations, identify limitations, and explain how they would improve the representation.
8	Contextual Interpretation of Fractions	This topic explores teachers' understanding of how context influences the meaning and operations of fractions. Teachers demonstrate awareness that the same mathematical operation can have different meanings and require different approaches.
9	Application of Fractions to Multi-Step Problems	This topic reveals teachers' approaches to solving complex fraction problems. Teachers demonstrate their ability to break down a problem involving nested fractions, showing multiple solution paths and explanations.
10	Critical Evaluation of Student Visual Representations	This topic centers on teachers' ability to interpret and evaluate student-created visual representations of mathematical concepts. Teachers demonstrate their skill in analyzing student work to identify understanding or misconceptions in mathematical representations.

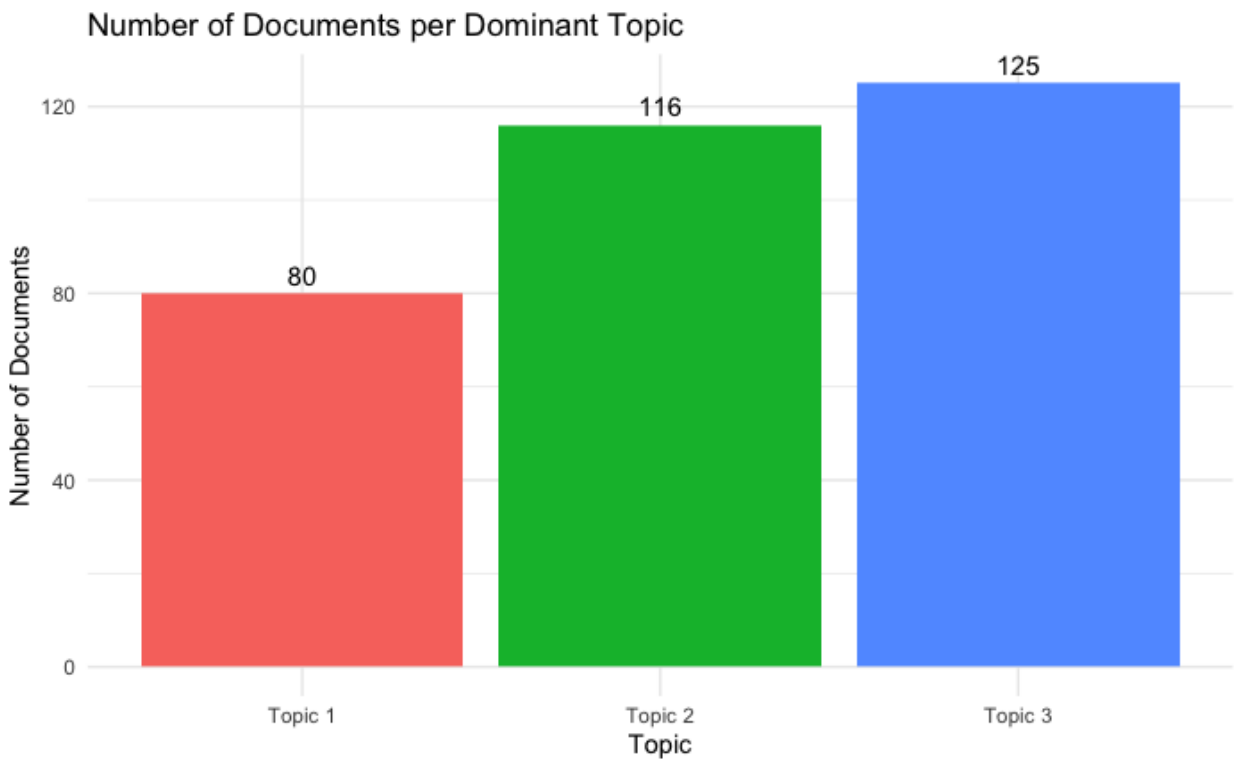

Figure 2
Model Selection

5.6. Data-Driven Validity Evidence

5.6.1. Evidence of Content Validity

The first source of content validity evidence is derived from the number and interpretation of the topics. The results in Table 5 and Figure 5 illustrate that a six-topic structure was coherent and interpretable, with each topic representing aspects of teachers' reasoning in fractions and proportions. This appears to satisfy the condition that the intended constructs should closely resample their topic definition.

Based on the proposed framework, each topic should ideally show a one-to-one relationship with an intended construct. However, the results indicate that most of the topics mapped onto more than one construct, except for Topic 1, which directly aligned with the Referent Unit. Topic 2 reflected Invariance and Covariance; Topic 3 aligned with Referent Unit and Invariance; Topic 4 with Invariance, Quantity, and Referent Unit; Topic

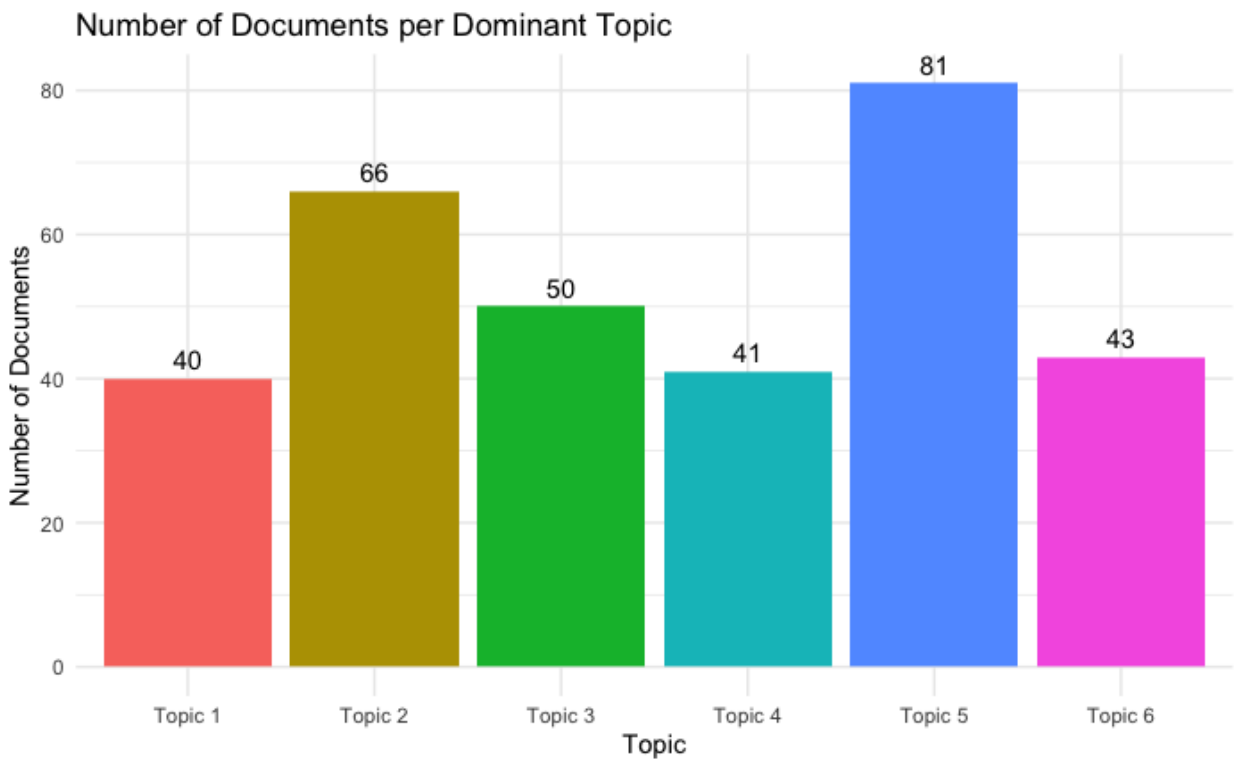

Figure 3
Model Selection

5 with Referent Unit and Quantity; and Topic 6 with Covariance and Referent Unit. These patterns suggest potential limitations regarding the extent to which constructs were exclusively reflected in teachers' responses.

The second source of content validity evidence comes from interpreting the likelihood of each item in each topic. The results in Table 7 show that although several items were designed to measure one construct, the latent topics derived from teachers' responses often reflected multiple constructs. For example, Items 1.3, 6.1, and 6.5 were simple items intended to measure Referent Units. However, answers to Item 1.3 mostly elicited words related to Topic 3, which closely aligned with Referent Unit and Invariance. Similarly, Items 6.1 to 6.5 primarily aligned with Topic 5, which included Referent Unit and Quantity words.

In some cases, items elicited responses related to a different construct than intended, indicating possible content misalignment or poor writing. For example, Items 7.1

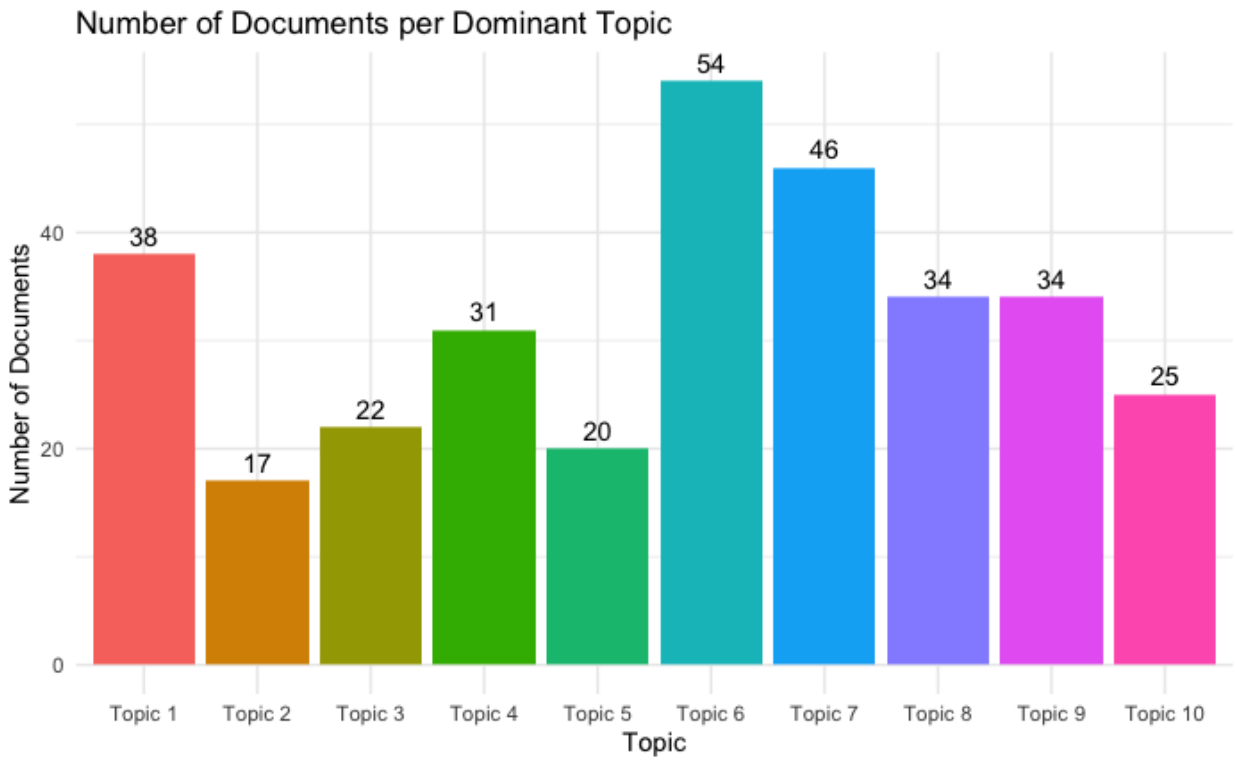

Figure 4
Model Selection

to 7.3 were intended to measure Quantity but elicited answers related to Referent Units and Quantity.

The third source of evidence concerns the items representativeness and balance across topics. As shown in Table 7, Referent Unit was the most frequently represented construct, aligned with its intended interview protocol. However, the empirical and intended distributions revealed gaps in content coverage, as most contents appeared embedded within multi-construct topics rather than being measured in isolation.

Table 7

Construct-to-Topic Alignment for Selected Items

Sub-item	Item Type	Construct(s) Measured	Highest Posterior Topic	Topic Alignment
1.3	Simple	Referent Unit	Topic 3	Referent Unit and Invariance
1.4	Complex	Quantity, Referent Unit	Topic 3	Referent Unit and Invariance
1.5	Complex	Quantity, Referent Unit	Topic 3	Referent Unit and Invariance
2.1	Complex	Invariance, Referent Unit	Topic 4	Invariance, Quantity, and Referent Unit
2.2	Simple	Invariance	Topic 4	Invariance, Quantity, and Referent Unit
4.1	Complex	Invariance, Quantity, Referent Unit	Topic 2	Invariance and Covariance
4.2	Complex	Invariance, Quantity, Referent Unit	Topic 2	Invariance and Covariance
4.3	Complex	Invariance, Quantity, Referent Unit	Topic 2	Invariance and Covariance
4.4	Simple	Referent Unit	Topic 2	Invariance and Covariance
6.1	Simple	Referent Unit	Topic 5	Referent Unit and Quantity
6.2	Simple	Referent Unit	Topic 5	Referent Unit and Quantity
6.3	Simple	Referent Unit	Topic 5	Referent Unit and Quantity
6.4	Simple	Referent Unit	Topic 5	Referent Unit and Quantity
6.5	Simple	Referent Unit	Topic 2	Invariance and Covariance
7.1	Simple	Quantity	Topic 1	Referent Unit
7.2	Simple	Quantity	Topic 1	Referent Unit
7.3	Simple	Quantity	Topic 1	Referent Unit
8.1	Complex	Covariance, Invariance, Referent Unit	Topic 6	Covariance and Referent Unit
8.2	Simple	Referent Unit	Topic 6	Covariance and Referent Unit
8.3	Complex	Covariance, Invariance	Topic 6	Covariance and Referent Unit
9.2	Complex	Quantity, Referent Unit	Topic 4	Invariance, Quantity, and Referent Unit
11.1	Simple	Referent Unit	Topic 5	Referent Unit and Quantity
11.2	Simple	Referent Unit	Topic 5	Referent Unit and Quantity

Figure 5

Sub-questions by Topics

5.6.2. Evidence of Internal Structure

Evidence of internal structure was evaluated using the posterior topic proportion and its corresponding credible interval based on item complexity and construct independence. Although topics were estimated as uncorrelated, many of them integrating multiple constructs.

Figure 6 illustrates the diagram between items and topics. Bold edges denote a strong and certain item-topic association, where the posterior and the lower bound of the credible interval are ≥ 0.7. Blue dashed edges typically represent complex items with high posterior topic proportion but wider intervals or mixed-topic associations. These edges indicate alignment with one topic but with some uncertainty around it. Red dotted edges reflect items with lower posterior mean (≤ 0.7) and wide intervals. These typically include complex items where one concept appears more salient than others.

Overall, the results showed that many simple items aligned well with their intended

constructs. For example, Item 1.3, which was designed to measure Referent Unit, had a posterior topic proportion of 0.91 on Topic 3 with a narrow 95% credible interval [0.87, 0.95], indicating strong and consistent alignment. Items 6.1 through 6.5, also simple and targeting Referent Unit, loaded on Topic 5 with proportions ranging from 0.74 to 0.87, providing further support for structural alignment. In contrast, Items 7.1 to 7.3, which were designed to measure Quantity, loaded on Topic 1 (Referent Unit). This suggests construct misalignment despite high posterior estimates.

Complex items showed more varied patterns. For instance, Item 4.1, which was designed to measure Invariance, Quantity, and Referent Unit, loaded primarily on Topic 2 (0.53; [0.43, 0.63]), suggesting that Invariance dominated participants' responses. While some complex items such as Item 8.1 (targeting Covariance, Invariance, and Referent Unit) loaded almost exclusively on a single multi-construct topic (Topic 6), this was interpreted as evidence of conceptual integration rather than imbalance.

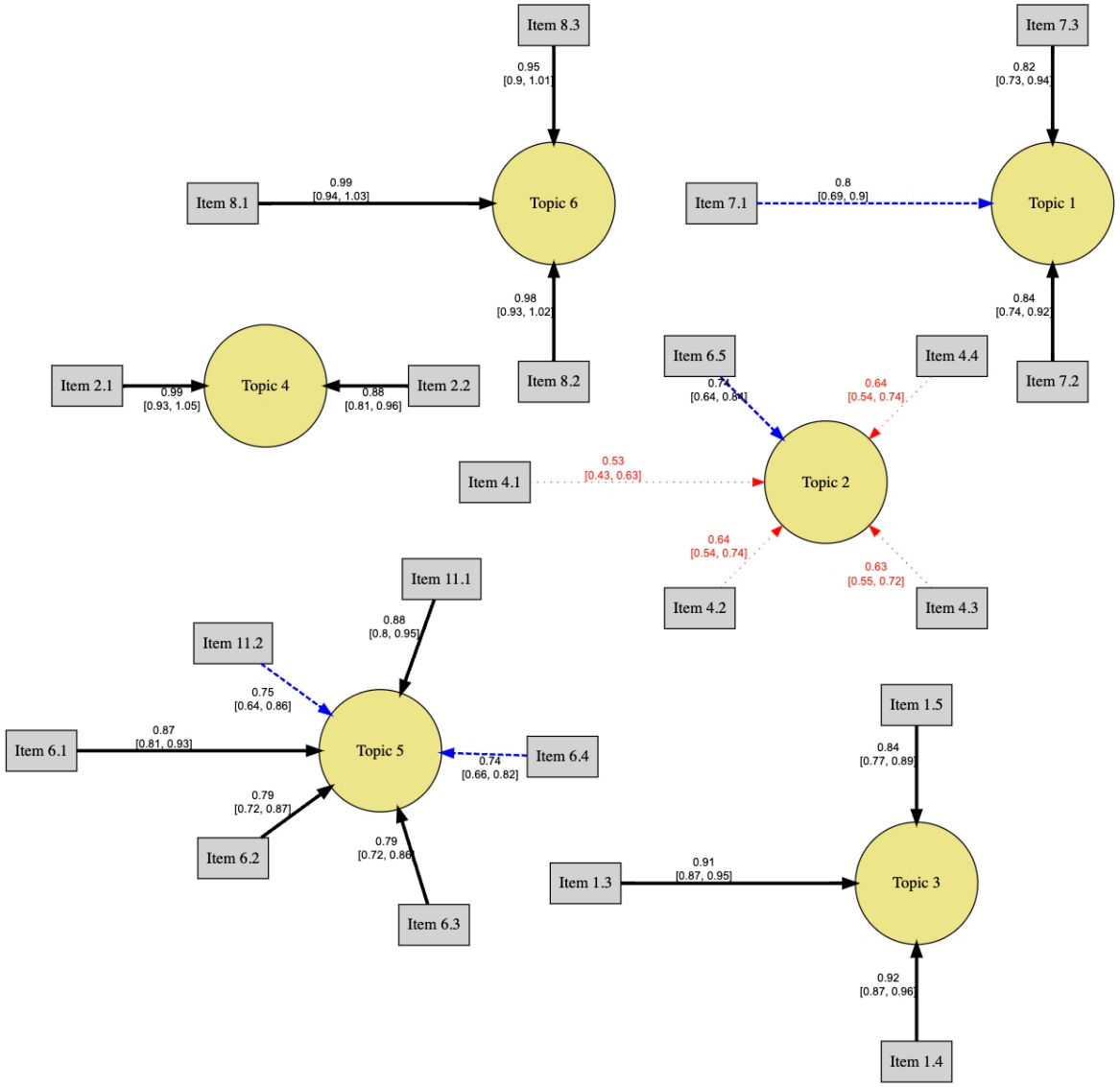

Figure 6

Item-Topic Diagram

6. Discussion and Conclusion

This study proposed a novel data-driven framework to provide validity evidence for text-based assessments and illustrated its uses through a corpus of teachers' written answers to an interview protocol designed to measure their knowledge of proportional reasoning and fractions before participating in a professional development intervention.

Although the resulting topic structure deviated from the ideal of clean construct separation, these results likely reflected the complexity of teacher reasoning in applied instructional contexts. In mathematics education, constructs such as Referent Unit and Quantity are conceptually related and often co-occur in practice. Moreover, many of the items designed for this instrument were intentionally complex, increasing the likelihood that the elicited responses would activate multiple constructs.

These findings may also reflect a form of hierarchical reasoning, in which teachers begin with foundational ideas about Referent Unit to then build upon concepts such as Quantity and Invariance. Thus, rather than being separated constructs, these ideas may emerge in a sequential fashion. Notably, Referent Unit was the most frequently intended construct to be measured by the instrument, which may suggest that this concept was not only pedagogically important but also implicitly recognized during the item development phase. Overall, the results supported, in part, the expected patterns based on item complexity and construct relationships, with narrow credible intervals providing evidence that the internal structure aligned with expectations.

To conclude, the proposed framework provides a systematic approach to gather evidence of content and internal structure in text-based assessment by fully capturing uncovered patterns that could have been lost when using traditional psychometric techniques. Its flexibility also supports simple and complex items, with or without topic relationships, offering an alternative approach to gather validity evidence. Furthermore, the research findings suggest that the proposed framework can be used in designing and refining textual tests, ensuring a more balanced and comprehensive distribution of items

across topics. However, given its data-driven nature, its results are grounded by the observed language patterns rather than from a pre-defined theoretical model. Therefore, while it allows for rich interpretations, the results are tied to a specific dataset.

References

Abraham, A., Mardones-Segovia, C., Sarles-Whittlesey, H., & Cohen, A. S. (2024). Themes and trends in creativity research between 1894 and 2022: A topic modeling approach. *Psychology of Aesthetics, Creativity, and the Arts*.

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in social and administrative pharmacy*, *15*(2), 214–221.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

American Educational Research Association, A., & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Bradshaw, L., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99–118.

Cardozo-Gaibisso, L., Kim, S., Buxton, C., & Cohen, A. (2019). Thinking beyond the score: Multidimensional analysis of student performance to inform the next generation of science assessments. *Journal of Research in Science Teaching*, *1*(57), 856–878.

Choi, H.-J., Kwak, M., Kim, S., Xiong, J., Cohen, A. S., & Bottge, B. A. (2019). An application of a topic model to two educational assessments. In *Quantitative psychology: 83rd annual meeting of the psychometric society, new york, ny 2018* (pp. 449–459).

Cifuentes, J., & Olarte, F. (2023). A macro perspective of the perceptions of the education

system via topic modelling analysis. *Multimedia Tools and Applications*, *82*(2), 1783–1820.

Dunwoodie, K., Macaulay, L., & Newman, A. (2023). Qualitative interviewing in the field of work and organisational psychology: Benefits, challenges and guidelines for researchers and reviewers. *Applied Psychology*, *72*(2), 863–889.

Guilford, J. P. (1957). Creative abilities in the arts. *Psychological review*, *64*(2), 110.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd edition)*. Internet: https://web. stanford. edu/~ jurafsky/slp3/,[Accessed: September 19, 2023].

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55–73.

Mardones-Segovia, C., Choi, H.-J., Hong, M., Wheeler, J. M., & Cohen, A. S. (2022). Comparison of estimation algorithms for latent dirichlet allocation. In *Quantitative psychology: The 86th annual meeting of the psychometric society, virtual, 2021* (pp. 27–37).

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Phakiti, A., & Isaacs, T. (2021). Classroom assessment and validity: Psychometric and edumetric approaches. *European Journal of Applied Linguistics and TEFL*, *10*(1), 3–24.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of statistical software*, *91*, 1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ...

Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064–1082.

Singh, D., & Singh, B. (2022). Feature wise normalization: An effective way of normalizing data. *Pattern Recognition*, *122*, 108307.

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema*, *35*(3), 217–226.

Sireci, S. G. (1998). The construct of content validity. *Social indicators research*, *45*, 83–117.

Spoto, A., Nucci, M., Prunetti, E., & Vicovaro, M. (2023). Improving content validity evaluation of assessment instruments through formal content validity analysis. *Psychological methods*.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Sun, J., & Yan, L. (2023). Using topic modeling to understand comments in student evaluations of teaching. *Discover Education*, *2*(1), 25.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.

Torlig, E., Junior, P. R., Fujihara, R., Demo, G., & Montezano, L. (2022). Validation proposal for qualitative research scripts (vali-quali). *Administração: Ensino e Pesquisa*, *23*(1).

Zumbo, B. D. (2006). 3 validity: Foundational issues and statistical methodology. *Handbook of statistics*, *26*, 45–79.