International Journal of Assessment Tools in Education
2025, Vol. 12, No. 4, 942–962
https://doi.org/10.21449/ijate.1653549
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Exploring trends in psychometrics literature through a structural topic model

Kübra Atalay Kabasakal[1*], **Duygu Koçak**[2], **Rabia Akcan**[3]

[1]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]Alanya Alaaddin Keykubat University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye
[3]Republic of Turkey Ministry of National Education, Afyonkarahisar, Türkiye

ARTICLE HISTORY
Received: Mar. 7, 2025
Accepted: July 8, 2025

Keywords:
Structural topic modelling,
Psychometrics,
Trend analysis,
Latent dirichlet allocation,
Text mining.

Abstract: The digitalization of knowledge has made it increasingly challenging to find and discover relevant information, leading to the development of computational tools to assist in organizing, searching, and comprehending vast amounts of information. In fields like psychometrics, which involve large datasets, a comprehensive examination of research trends, as well as understanding the prominence of various themes and their evolution over time through these tools, is essential for assessing the dynamic structure of the field. This study aims to explore the themes addressed in publications from eleven leading journals in psychometrics and to determine the overall distribution of topics. To achieve this, structural topic modelling has been employed. A comprehensive analysis of 8,523 article abstracts sourced from the Web of Science database revealed the existence of fourteen topics within the publications. "Scale Development and Validation" emerged as the most prominent topic, whereas "Differential Item Functioning" was the least well-known. The distribution of topics across academic journals emphasized the key role journals play in shaping the development and evolution of psychometric research. Through further exploration of topic correlations, potential future research directions and between-topic research areas were revealed. This study serves as a valuable resource for researchers aiming to keep up with the latest advancements in psychometrics. The findings provide crucial insights to guide and shape future research in the field.

1. INTRODUCTION

Psychometrics, although a field that has significantly advanced since the 2000s (Groenen & Ark, 2006), has much deeper historical roots. The foundations of psychometrics were established in the late 19th century by Sir Francis Galton, who aimed to evaluate human abilities using statistical methods and measurement techniques (Michell, 2022). In the post-World War II era, the emergence of psychometric methodologies and their applications in various domains contributed to noteworthy progress in the discipline (Jones & Thissen, 2006). In the mid-20th century, Charles Spearman introduced the concept of general intelligence, referred to as the "g

*CONTACT: Kübra ATALAY KABASAKAL katalay@hacettepe.edu.tr Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456

factor," and developed factor analysis, a fundamental method for identifying the common factors underlying psychological tests (Buckhalt, 1999).

This trajectory highlights how psychometric research in the 2000s gained momentum in response to the growing demands for psychological and educational measurement, paralleled by advancements in statistical and computational methods. During this period, particular emphasis was placed on the development and evaluation of psychometric tools, with researchers striving to ensure their reliability, validity, and applicability across diverse contexts (Martin & Savage-McGlynn, 2013). Simultaneously, interest in psychometric theory was revitalized, prompting academics to revisit foundational concepts and explore novel approaches to measurement and scale development (Jones & Thissen, 2006). This renewed focus drew attention to topics such as the psychometric validity of scales, item bias, differential item functioning, and estimation methods based on item response theory.

As researchers sought to bridge the gap between academic studies and practical applications, interest in the usability and interpretability of psychometric scales also increased (Vitoratou & Pickles, 2017). These advancements have enabled researchers to address increasingly complex and multifaceted research questions, thereby enhancing the depth and sophistication of psychometric analyses (Blanca *et al.*, 2018). By the 21st century, psychometric practices had become more sophisticated through the integration of technologies such as computer-assisted testing and data analytics. Psychometricians have utilized technological advancements to address emerging demands and develop the discipline. For example, the early 20th century witnessed a growing demand for standardized tests in education, marking a turning point in the development of measurement tools. During this period, pioneers like Thorndike emphasized the importance of measurement and evaluation practices in education and worked to establish a scientific foundation for these applications. The development and widespread adoption of educational achievement tests contributed significantly to the theoretical and practical growth of psychometrics.

Technological advancements, the growing demand for more sophisticated measurement tools, and an increasing emphasis on fairness and equity in psychological and educational assessment has shaped this progress in the field. For instance, innovations such as cognitive diagnostic modeling and adaptive testing have expanded the discipline's scope. However, the rise of large textual datasets has made organizing and understanding the themes in literature increasingly complex. Understanding the prominence of different themes in psychometric literature, how these themes have evolved over time, and how they vary across journals is crucial for evaluating the dynamic structure of the field. The interdisciplinary nature of psychometrics and its broad application in education, healthcare, and business further underscores the necessity of such an analysis. In this respect, topic modeling methods could be a desirable alternative for uncovering the hidden themes and trends in such a broad field.

Topic modeling, a powerful text mining technique that has become popular in natural language processing, can provide valuable insights into the themes and trends emerging in psychometrics literature (Gao & Sazara, 2023). This method's main goal is to identify underlying themes or topics within a large text corpus without prior content knowledge or labeling. One of the greatest advantages of topic modeling is its ability to process unstructured text data, which is ubiquitous in the digital age (Blei, 2012). The versatility of this technique makes it an essential tool for researchers and practitioners across a wide range of disciplines. Furthermore, its ability to handle diverse textual data, from short-form content like social media posts to long-form academic articles, underscores its adaptability (Richardson *et al*., 2014). Boon-Itt and Skunkan (2020), for instance, leveraged topic modeling and sentiment analysis to understand public awareness of COVID-19 trends and identify significant themes of concern shared by Twitter users. Polatgil (2023) employed topic modeling to analyze user comments on the Duolingo mobile app, a widely utilized tool among language learners, with the aim of identifying the key aspects highlighted by users. Another study by Hwang *et al.* (2023) used topic modeling

techniques to identify research trends in published articles on the use of technology in mathematics education.

The application of topic modeling is increasingly gaining prominence in educational measurement. Anderson *et al.* (2020) introduced a novel approach to gather content-related validity evidence, incorporating topic modeling as a key method. Wheeler *et al.* (2024) stated that topic modeling is becoming more widespread in educational measurement research, particularly for analyzing responses to constructed-response items. A recent study by Xiong and Li (2023) employed topic modeling methods in the development of automatic scoring algorithms for constructed-response items. The growing use of topic modeling for various purposes in the literature, along with the need for a comprehensive examination of research trends in fields like psychometrics that involve large datasets, has led to the emergence of this study.

Despite the increasing use of topic modeling in educational and psychological research, there has been no comprehensive investigation of its application to psychometric literature. This presents a gap in understanding how core themes have evolved within the field and how they differ across publication venues. The lack of such analysis limits our ability to grasp the intellectual structure and thematic development of psychometrics as an interdisciplinary domain. Therefore, the present study is significant in that it systematically maps the landscape of psychometric research using structural topic modeling (STM), providing insights into its conceptual evolution and the distribution of topics across journals and time. Specifically, this study seeks to address the following questions:

(1) What are the prominent themes in psychometric research?
(2) How have these themes evolved over time?
(3) How are these themes distributed across different journals?

In recent years, widespread adoption of technologies such as big data analytics and machine learning in psychometrics has significantly enhanced the field's capacity to perform complex analyses on large datasets. These technological advancements have also improved the validity and reliability of psychometric tools, enabling researchers to tackle increasingly complex challenges. Furthermore, the thematic organization and analysis of extensive literature have become crucial for guiding more focused and meaningful progress in the field. Such analyses not only provide valuable insights into the structure and focus of past research but also offer a framework for understanding future research orientations. To our knowledge, a comprehensive topic modeling and bibliometric analysis study in psychometrics remains absent. The only related example in the literature is the bibliometric investigation conducted by Zagaria and Lombardi (2024), which utilized the PsycINFO database to examine the relative prominence of Bayesian and frequentist approaches in the fields of psychology and psychometrics.

This study aims to explore the role of psychometrics in the scientific world and identify key focus areas for the future. It uses Structural Topic Modeling to analyze prominent themes in psychometric literature, their development over time, and their distribution across journals. This analysis provides a framework for understanding how fundamental psychometric methodologies have evolved and which themes have gained prominence in response to societal and technological changes. The thematic analysis further sheds light on the interdisciplinary nature of the field, highlighting the relationship between its theoretical foundations and practical applications. Together, these insights offer a more comprehensive understanding of the evolving landscape of psychometrics and its response to emerging demands and opportunities.

1.1. Structural Topic Modeling

Topic modeling is a machine learning approach that uses probabilistic models to uncover the themes and semantic patterns in large volumes of unstructured text data. By analyzing and linking documents based on word frequency patterns, these models can identify a set of

"topics," with each word and document associated with one or more topics (Blei, 2012). In topic modeling, each topic is defined by a collection of semantically related words. The model identifies these relationships by examining the frequency of words across the entire corpus of text. This enables a single word to relate to multiple topics, as its usage can vary depending on the context. Instead of assigning a single topic to each document, the model produces a probability distribution indicating the likelihood of a document belonging to each identified topic, based on the word patterns present (Blei *et al.*, 2003; Blei, 2012).

Structural topic modeling (STM) builds upon standard topic modeling by incorporating document metadata into the topic prediction process. This approach differs from traditional classification methods, which typically assign a document to a single, discrete category. STM is grounded in Latent Dirichlet Allocation (LDA; Blei *et al.*, 2003), but with a key distinction: LDA assumes topic prevalence and word usage patterns are static across all documents, while STM accounts for variability in these patterns by allowing them to be influenced by relevant covariates (Tonidandel *et al.*, 2021).

STM is a valuable tool for analyzing large volumes of unclassified text data. This advanced modeling approach identifies meaningful linguistic and semantic connections within the text and uncovers how these patterns vary across relevant metadata. STM's ability to link words with similar meanings and to distinguish multiple usages of the same word provides a more nuanced and contextual understanding of the content, making it particularly useful for analyzing extensive, unstructured datasets. By incorporating document metadata, STM enables researchers to examine the relationships between topic content, topic prevalence, and external variables (Tonidandel *et al.*, 2021). This technique can uncover hidden themes and trends in large text corpora and provide insights that may be missed by other methods.

Overall, STM is a powerful and versatile analytical framework for extracting meaningful insights from large, unstructured text datasets across a variety of contexts, including open-ended survey responses, news articles, and social media posts. This makes it an invaluable asset for researchers in fields such as social sciences, business, education, and public health (Bai *et al.*, 2021; Roberts *et al.*, 2014).

In the context of this study, STM offers a robust methodological framework for identifying the key themes within psychometric literature and examining how these themes evolve across time and publication venues. This approach is particularly valuable given the increasing volume and complexity of research in psychometrics, where traditional content analysis methods may fall short. By incorporating document-level metadata such as publication year and journal, STM enables a more nuanced exploration of thematic shifts in the field. Therefore, this study applies STM not only to map the conceptual landscape of psychometrics but also to uncover the dynamic interplay between methodological developments and thematic emphasis. Through this, we aim to contribute a systematic and scalable method for organizing knowledge in psychometrics and guiding future research directions.

2. METHOD

This study employed a descriptive and exploratory research design based on structural topic modeling (STM). The purpose was to examine the thematic structure of psychometric research by identifying latent topics within a large body of scholarly literature and exploring their temporal and journal-based variations. The following subsections outline the data collection, preprocessing, and analytical procedures used in the study.

2.1. Data Collection

We initiated our analysis of psychometric literature by identifying the most relevant journals in this field. Utilizing bibliometric techniques, we selected journals with a high volume of articles focused on psychometrics for further examination. To investigate the psychometrics publication landscape, we first identified the eleven most relevant journals in the field. After a

comprehensive review of the databases indexing these journals, we determined that Web of Science (WoS) was a primary source of bibliographic information. In the WoS search, we considered journals in both SSCI and ESCI indexes.

We retrieved the data on 26 January 2025 and extracted the publication name, publication year, journal name, document type, and abstract metadata from the WoS database and downloaded 1000 articles each time. We then examined the downloaded files to ensure they were from journal sources, the document type was an article, and an abstract was present in the metadata. After this examination, we consolidated the data into a single file. Our literature search identified 19,826 publications, but 10,829 of these lacked abstracts. This resulted in 8,997 publications remaining. An analysis of the publication years for the abstracted publications revealed that 46 were published before 1989, and these 46 were excluded from the corpus. Finally, the document type was restricted to articles, and this process left 8,523 articles. Information regarding 8,523 articles comprising the corpus is presented in Table 1. The names and abbreviations of the academic journals, the number of articles from each journal in the corpus, and the publication year ranges of the articles are presented in Table 1.

Table 1. *Features of the articles in the corpus.*

Journal Name and Abbrev	*n*	Publication Year Range
Applied Measurement in Education (AME)	549	1995- 2024
Applied Psychological Measurement (APM)	1009	1991- 2025
Educational and Psychological Measurement (EPM)	2146	1991- 2025
Educational Assessment (EA)	289	2005- 2024
Educational Measurement-Issues and Practice (EM-IP)	315	2013- 2024
International Journal of Assessment Tools in Education (IJATE)	395	2014- 2024
Journal of Educational and Behavioral Statistics (JEBS)	712	1994- 2025
Journal of Educational Measurement (JEM)	698	1992- 2025
Journal of Measurement and Evaluation in Education and Psychology (EPOD)	307	2010- 2024
Psychometrika (PSYCH)	1321	1990- 2024
Studies in Educational Evaluation (SEV)	782	2013- 2025

As shown in Table 1, most articles in the corpus are published in EPM, while the fewest originate from EA. Another noteworthy detail in Table 1 is that the EM-IP, IJATE, and EPOD journals, which show lower representations, are recent publications.

2.1.1. *Data preparation*

Before applying topic modeling, several text preprocessing steps were required. The text was converted to lowercase, and punctuation, numbers, and symbols were removed. Additionally, stop words (frequently used words) that provide little semantic meaning, such as "the", "a", "by", and "so", were eliminated. This is a widespread practice, as stop words appear frequently in text yet contribute little to the analysis. There is no universally accepted dictionary of stop words, but various libraries are available. For this application, the stop word lexicon from the tidytext R package was utilized (Silge & Robinson, 2016). We also removed terms like "approach", "data", "research", "article", "set", "procedure", and "framework" as sources of noise in the text, which led to the emergence of the Research Methods and Data Analysis topic. This text preprocessing step helped improve the performance of the language classification algorithm, such as STM, by focusing the analysis on more substantive and informative content within the text documents (Banks *et al.*, 2018).

Common text normalization techniques include stemming and lemmatization (Banks *et al.*, 2018). These methods reduce words to their roots but differ in their approaches. Stemming uses

pattern-based methods to determine the root without considering vocabulary, context, or parts of speech. In contrast, lemmatization involves morphological analysis to extract the root word while preserving semantic meaning (Singh & Gupta, 2017). Previous research has yielded mixed findings on the added benefits of applying stemming or lemmatization to large text datasets (Banks *et al.*, 2018). For this analysis, the lemmatization approach is selected as it maintains the conceptual significance of the words.

The article abstracts were pre-processed and converted into a traditional word-document matrix. In this matrix, each row represents the text from an individual abstract, and each column denotes a word used in the entire text corpus. In addition to single words, we also incorporated two-word combinations (bigrams) in our analyses. The decision to include unigrams, bigrams, and even trigrams reflects the need to balance adequately capturing meaning without unnecessarily increasing model complexity. As the n-gram size increases, the number of columns in the dataset grows exponentially, resulting in an extremely sparse, high-dimensional data structure. Unfortunately, high-dimensional data presents various analytical challenges that can impede the application of numerous techniques, a phenomenon known as the curse of dimensionality. While multiple n-grams should be avoided, they may be desirable if there is reason to believe they convey important semantic meaning beyond single words. In our case, we made an a priori decision to include bigrams because we believed that two-word combinations, such as "formative assessment," "rater reliability," or "internal consistency," could reflect nuanced conceptual relationships. To mitigate the impact of sparse words, which contribute little to understanding common topics but can increase the computational complexity of structural topic models, we used a lower inclusion threshold. Only words or bigrams appearing in more than five documents were retained. Our final text corpus consisted of 8,523 documents, 13116 terms, and 384004 tokens.

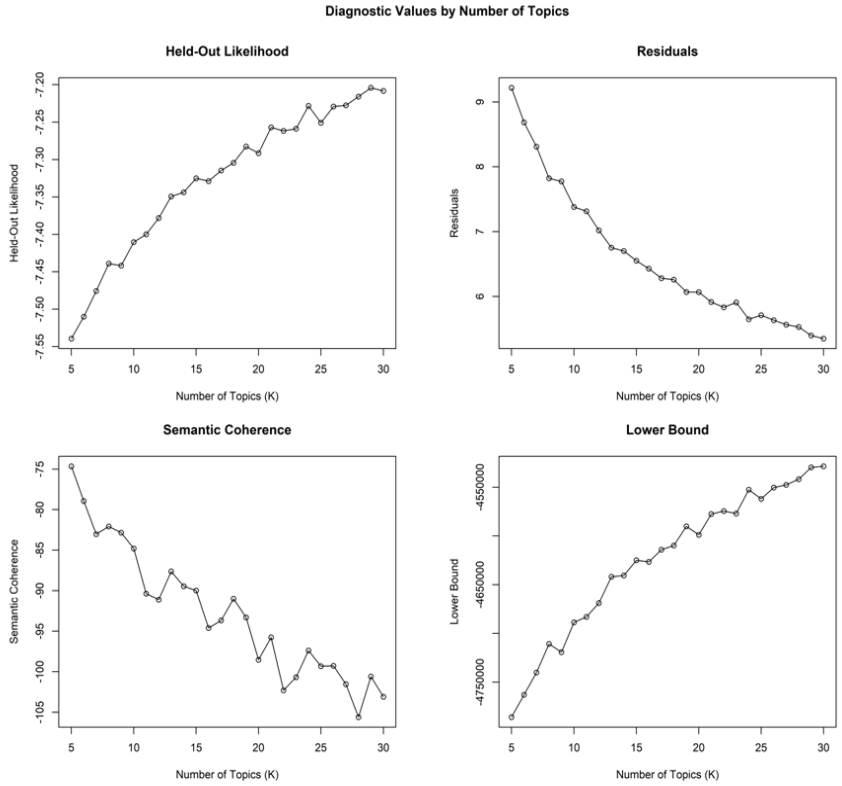
2.1.2. *Data analysis*

The STM analysis was conducted using the stm package (Roberts *et al.*, 2019) in R software. The number of topics (K) was determined by comparing models using semantic coherence and exclusivity metrics. Document metadata, including publication year and journal name, were incorporated as covariates to examine their influence on topic prevalence. The model output includes topic-word distributions (β), document-topic distributions (θ), and estimates of topic variation over time and across journals.

To determine the optimal number of topics, we used an iterative approach with the searchK() function in stm, estimating models with 5 to 30 topics. This process is analogous to inspecting a scree plot in exploratory factor analysis (Tonidandel *et al.*, 2021). As shown in Figure 1, key metrics guide topic selection. Semantic coherence measures how frequently a topic's most probable terms co-occur, held-out likelihood assesses predictive performance on unseen data, residual variance indicates unexplained variation, and the lower bound reflects model log-likelihood, with higher values signifying better fit.

When determining the number of topics, the goal is to strike the best balance between coherence, exclusivity, and cohesion. As more topics are added, exclusivity tends to increase, enabling more refined differentiation. However, this also tends to reduce semantic coherence. The sweet spot is where coherence remains decent, but exclusivity is high. Based on our analysis, the models with 8 to 15 topics appeared to strike this balance. After reviewing these options, the 14-topic model was identified as the most meaningful and interpretable.

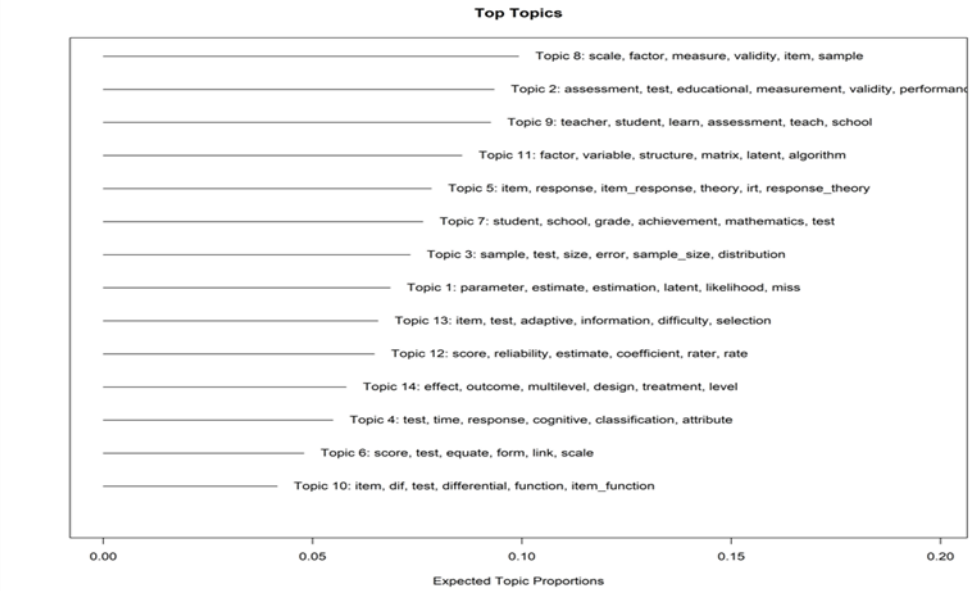
Figure 1. *Comparison of solutions for topics spanning 5–30.*



3. RESULTS

The analysis of 14 topics revealed their relative prevalence within the overall corpus, as depicted in Figure 2. This figure also presents the six most frequent words associated with each topic. Topic 8, characterized by the terms factor, scale, and measure, emerged as the most dominant topic. Conversely, Topic 10, containing the words item, differential, test, and function, demonstrated the lowest prevalence. The topic prevalence values span a range from 5% to 10%, with the four most dominant topics occupying the upper end of this spectrum, while the remaining topics have a prevalence of approximately 5% in the corpus, as can be seen in Figure 2.

Figure 2. *Topics by prevalence.*



STM uses metrics like prob (highest probability), FREX (frequency-exclusivity), lift, and score to uncover thematic patterns. Prob indicates the likelihood of a word belonging to a topic, while lift measures its uniqueness. FREX balances frequency and exclusivity, identifying words that are both common and distinctive (Roberts *et al.*, 2014). Score evaluates topic prevalence and distinctiveness, providing a comprehensive assessment of importance. In this study, topic labelling relied on high-probability words and FREX metrics, supplemented by expert review of the top five articles for each topic. Word clouds (Figure 3) visually represent the most probable words for each topic, with font size indicating probability.

Figure 3. *Word clouds for each topic.*

Word clouds illustrate the words with the highest probability of occurrence within each topic, where the font size corresponds to the probability of the word appearing. Figure 3 indicates that each topic is characterized by a unique set of related words. Furthermore, the most relevant words based on the FREX metric, along with the five most representative article abstracts for each topic, were examined in terms of context and depth of thematic content. The topics were then labeled based on the highest probability of word occurrence and the FREX metric. The topic naming process considered these metrics in conjunction with the content of the related articles, and the details are explained below.

The most prevalent topic identified in the analysis was Topic 8, Scale Development and Validation. This naming was derived from the five most representative article abstracts associated with this topic, which predominantly focus on the development, validation, and psychometric evaluation of various scales. For instance, Göral *et al.* (2024) conducted a methodological study to adapt and validate the Attitudes to Fertility and Childbearing Scale in a Turkish context, demonstrating strong reliability and validity through confirmatory factor analysis and internal consistency measures. Similarly, Tharenou and Terry (1998) assessed the reliability and validity of subjective and behavioral measures of managerial aspirations,

highlighting their distinct but related constructs and satisfactory psychometric properties. In another study, Tunç *et al.* (2021) developed the Hostility in Pandemic Scale to measure hostility levels during the COVID-19 pandemic, confirming its one-dimensional structure and high reliability through exploratory and confirmatory factor analyses. Lastly, Gregson (1991) explored the relationship between communication satisfaction and job satisfaction, using factor analysis to establish their separability as constructs. Collectively, these studies underscore the centrality of scale development and validation in psychometric research, as reflected in Topic 8.

The second most prevalent topic was Topic 2. Based on the FREX metrics and insights from articles in the corpus, this topic was titled as follows: Validity and Equity in Educational and Psychological Testing. This topic includes studies that examine the revision process and changes in the Standards for Educational and Psychological Testing, updated in 2014 (Plake and Wise, 2014), as well as research exploring how test results are evaluated as validity evidence (Cizek *et al.*, 2010). Additionally, discussions focus on how validity theory is shaped by the contexts in which tests are used (Sireci, 2013) and how core competencies in educational measurement can be developed (Ackerman, 2023). Furthermore, this topic emphasizes the need to consider assessment processes within a framework of social responsibility and justice (Buzick *et al.*, 2023) and discusses how the concept of validity can be expanded from a racial justice perspective (Lederman, 2023; Randall *et al.*, 2022). Studies in this topic also address the design and quality control processes of automated scoring systems (Rupp, 2018) and provide recommendations for the role of artificial intelligence in educational measurement, ensuring its use aligns with ethical standards (Briggs, 2024).

The third most prevalent topic was Topic 9 (Formative Assessment and Feedback). Brooks *et al.* (2020) evaluated the impact of a professional learning intervention using a student-centred feedback model in primary schools. Another study by Bastola and Hu (2021) examined students' views on feedback from thesis supervisors at a Nepalese university. Students felt the feedback was inadequate, but still engaged with it. The study also found that feedback engagement varied by discipline, suggesting the need for subject-specific feedback practices. Jiang and Ironsi (2024) explored how students respond to corrective peer feedback in the classroom. The results revealed that while students saw peer feedback as helpful, they also felt it was sometimes unfair or improperly assessed. These studies may demonstrate that the need for research on feedback in different fields and from various perspectives has led to this outcome.

The fourth most prevalent topic was Topic 5, Item Response Theory. This topic encompasses research focused on improving and refining IRT models, particularly concerning the analysis, scoring, and interpretation of data in educational testing and psychometrics. Studies such as those by Cohn & Huggins-Manley (2019), Huynh (1996), and Van Der Ark (2005) have contributed to advancements in these areas, emphasizing the development of more precise and reliable measurement techniques.

Topic 11, Structural Equation Modelling, represents another significant research focus on the corpus. Studies within this topic explore advanced methodologies in multivariate statistical analysis, particularly dimensionality reduction techniques, optimization algorithms, and methods for analysing complex, multi-dimensional data structures. Researchers such as Choi *et al.* (2016) and Kiers (1997) have contributed to the evolution of these techniques, aiming to develop more efficient ways to extract meaningful information from large datasets.

Topic 7, Exploring Predictors of Academic Performance, includes studies that investigate several factors influencing academic success. Some research explores the relationship between standardized test scores (e.g., SAT, GRE, GMAT) and academic performance in higher education (Meeter, 2022). On the other hand, some studies examine how socioeconomic status,

gender, and ethnicity affect student achievement at different educational levels, from primary school to university (Yavuz *et al.*, 2016).

Topic 3, Simulation-based Power and Error Analysis, includes studies that aim to improve and evaluate statistical methods used in hypothesis testing, multiple comparisons, and model evaluation. This research provides better tools and guidelines for conducting robust statistical analyses across different conditions and data types. For instance, MacDonald and Gardner (2000) used Monte Carlo methods to assess Type I error rates of six post hoc tests under various conditions. In a related approach, Guo and Luh (2008) proposed a method for determining appropriate sample sizes for Welch's F test in the presence of unequal variances through Monte Carlo simulations.

Topic 1, Parameter Estimation with Bayesian Methods, consists of studies discussing various methods for estimating item parameters in IRT models, including marginal Bayesian estimation, maximum likelihood estimation, and Gibbs sampling. These methodological advancements enhance the accuracy and applicability of IRT in psychometric research. In this context, Kim (2001, 2006) conducted studies on estimating item parameters in IRT models, particularly comparing calibration methods and evaluating the specific performance of MCMC-based Gibbs sampling for item and person parameter estimation.

Topic 13, Adaptive Testing and Test Assembly, focuses on advanced research in computerized adaptive testing (CAT). The studies explore methods for item selection, exposure control, and content balancing, aiming to enhance the efficiency, security, and fairness of adaptive tests while maintaining measurement precision (Chen & Lei, 2005; Han, 2012; Pan *et al.*, 2023).

Topic 12, Rater Reliability and Generalizability, encompasses research efforts to improve and evaluate various reliability and agreement measures used in psychometrics. Studies in this topic provide insights into the properties, limitations, and appropriate applications of these measures in different psychological and educational contexts. Some studies critique coefficient alpha for its limitations in handling correlated errors (Sijtsma & Pfadt, 2021; Rae, 2006), while another study discusses the relationships between weighted kappa, intraclass correlation, and product-moment correlation to better understand differences in interpretation (Schuster, 2004).

Topic 14, Mediation Effects and Multilevel Modeling, represents research in causal inference, focusing on methods for estimating causal effects in complex research designs (Park *et al.*, 2018; Talloen *et al.*, 2016). This topic addresses challenges such as confounding, noncompliance, and heterogeneity in treatment effects while providing tools for more robust causal inferences in fields such as education, social sciences, and medical research.

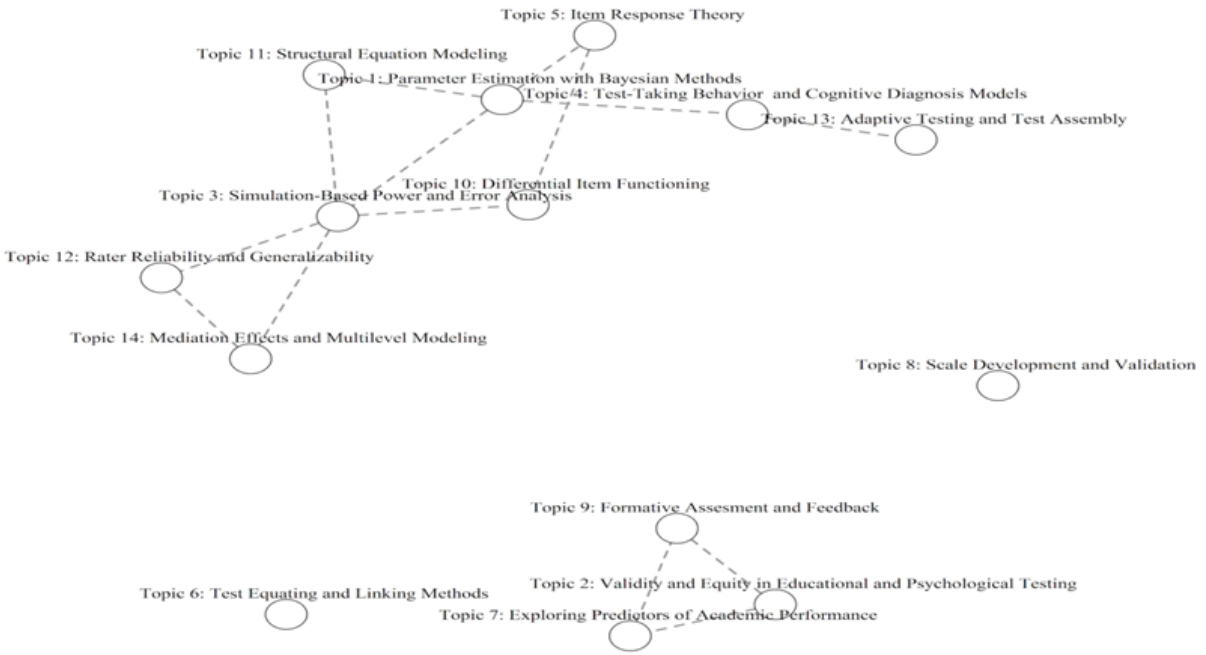
Topic 4, Test-Taking Behaviour and Cognitive Diagnosis Models, includes research on cognitive diagnostic assessment and modelling. Studies explore ways to improve the accuracy and interpretability of diagnostic information obtained from educational and psychological tests, tackling issues such as attribute specification, response time modelling, and integrating multiple data sources to enhance assessment precision. To give an example, a study by Zhan *et al.* (2022) proposes a multimodal joint cognitive diagnosis model that integrates accuracy, response times, and visual fixation counts from eye-tracking data. They also used an empirical example to demonstrate the applicability and benefits of the proposed model.

Topic 6, Test Equating and Linking Methods, focuses on comparing and refining test equating methods, understanding their theoretical foundations, and evaluating their performance under different conditions. This research aims to enhance the accuracy and reliability of test score comparisons across different forms and populations, ensuring fairer and more valid assessments. Numerous studies in the corpus examine the effectiveness of a wide range of equating techniques, often focusing on specific equating designs (Jiang *et al.*, 2012; Liu & Low, 2008; Von Davier *et al.*, 2004). Additionally, researchers also investigate the impact of sample size, population differences, and the type of anchor test used in the equating process (Liu & Low, 2008; Skaggs, 2005).

Topic 10, Differential Item Functioning (DIF), represents advanced research focused on improving the accuracy, power, and interpretability of DIF analyses in psychometric testing. The research aims to enhance the fairness and validity of educational and psychological assessments across distinct groups of test-takers. To this end, several studies have been conducted to assess the performance of common DIF detection techniques across different datasets and conditions and compare traditional methods with newer approaches (J. Chen *et al.*, 2013; Hidalgo & LÓPez-Pina, 2004; Shih & Wang, 2009).

The STM analysis enables the identification of correlated topics by analyzing their co-occurrence patterns within the same documents. As depicted in Figure 4, the topic network illustrates connections between nodes (topics) that exhibit a high probability of co-occurrence. Specifically, an edge is drawn between two nodes if their correlation coefficient exceeds 0.02. The observed topic correlations often reflect intuitive and meaningful relationships between the underlying concepts. The topic correlation analysis suggests that the psychometrics domain encompasses a distinct and diverse set of topics. For instance, Topic 8 (Scale Development and Validation) and Topic 6 (Test Equating and Linking Methods) were positioned separately from the other topics. Topic 2 (Validity and Equity in Educational and Psychological Testing), however, was linked to both Topic 7 (Exploring Predictors of Academic Performance) and Topic 9 (Formative Assessment and Feedback). Topic 1 (Parameter Estimation with Bayesian Methods) seemed related to several topics, including Topic 3 (Simulation-based Power and Error Analysis), Topic 4 (Test-taking Behavior and Cognitive Diagnosis Models), Topic 5 (Item Response Theory), and Topic 11 (Structural Equation Modeling). Topic 1 here can be regarded as a key concept that connects these topics due to its wide-ranging applications in psychometrics. Additionally, Topic 12 (Rater Reliability and Generalizability) and Topic 14 (Mediation Effects and Multilevel Modeling) were correlated, and they both were connected to Topic 3 (Simulation-based Power and Error Analysis). The graph also indicated that Topic 10 (Differential Item Functioning) shared connections with Topic 3 (Simulation-based Power and Error Analysis) and Topic 5 (Item Response Theory). This result is expected, as these three methods can be employed together to enhance test fairness and reliability.

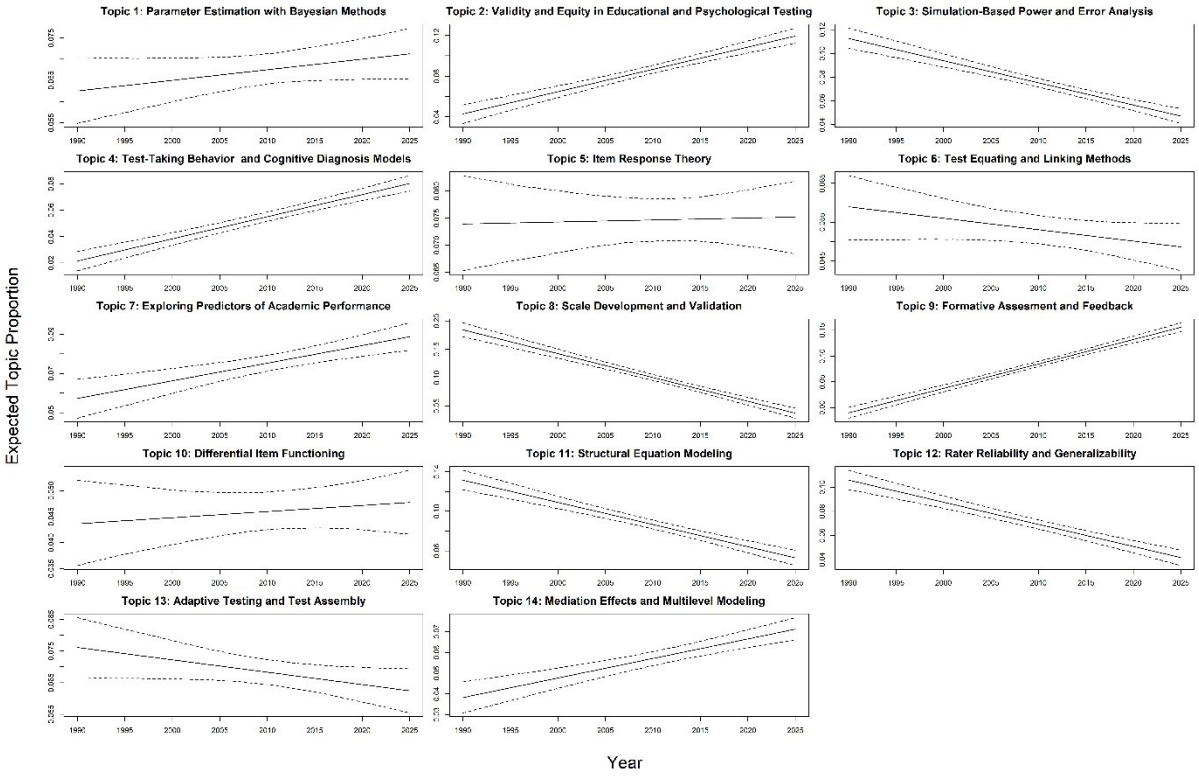
Figure 4. *Network of topic correlation.*



In the application of STM, the choice of covariates to include in the model is a crucial decision. In this study, the researchers selected the year of publication and journal names as covariates that may influence topic popularity. The key advantage of STM is its ability to investigate the

interactions between these covariates and the identified topics. By incorporating the publication year as a covariate, the model can track the prevalence of topics over time and enable comparative analysis. The topic prevalence is estimated as a function of the publication year, and confidence intervals around the estimated topic proportions are also generated. The temporal dynamics of topic popularity are presented in Figure 5, which illustrates the changes in the prevalence of each topic.

Figure 5. *Expected topic proportions over time with %95 confidence intervals for each topic.*



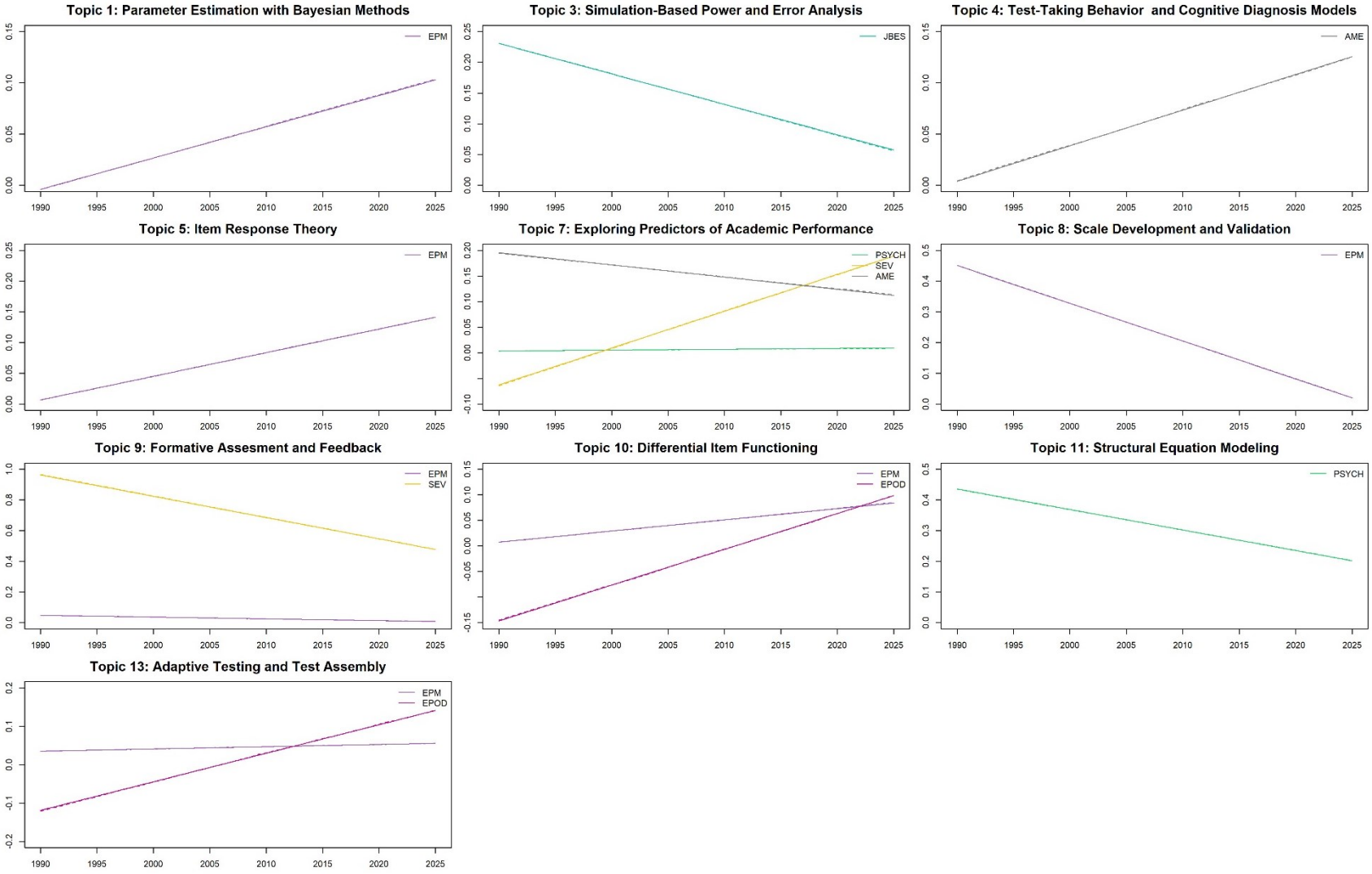
As shown in Figure 5, the researchers identified several "hot" and "cold" topics, reflecting those with increasing and decreasing prevalence trends, respectively, over the past 35 years. The "hot" topics include: Topic 2 (*Validity and Equity in Educational and Psychological Testing*), Topic 4 (*Test-Taking Behavior and Cognitive Diagnosis Models*), Topic 7 (*Exploring Predictors of Academic Performance)*, Topic 9 (*Formative Assessment and Feedback*), and Topic 14 (*Mediation Effects and Multilevel Modeling*). These topics indicate growing interest and relevance in the field. In contrast, the "cold" topics include: Topic 3 *(Simulation-Based Power and Error Analysis)*, Topic 8 *(Scale Development and Validation),* Topic 11 *(Structural Equation Modeling)*, Topic 12 *(Rater Reliability and Generalizability),* and Topic 13 (*Adaptive Testing and Test Assembly)*. These trends suggest a decline in research focus or interest in these areas over time. Additionally, several topics showed no significant relationship with the year of publication, indicating stable or inconsistent trends in their prevalence. These topics include: Topic 1 *(Parameter Estimation with Bayesian Methods)*, Topic 5 *(Item Response Theory*), Topic 6 *(Test Equating and Linking Methods)*, and Topic 10 *(Differential Item Functioning)*. The lack of significant trends in these areas may reflect consistent but unchanging interest or methodological stability over the years.

After STM analysis, a regression model can be constructed with each document serving as the unit of analysis. The dependent variable is the proportion of each document associated with a particular topic in the STM model, while the independent variables are the document metadata. This approach enables examination of the main and interaction effects of the covariates after the STM analyses are conducted. The publication year variable considered a covariate in this study exhibited a significant effect on the topic proportions in all topics except for Topic 1

(*Parameter Estimation with Bayesian Methods*), Topic 5 (*Item Response Theory*), Topic 6 (*Test Equating and Linking Methods)*, Topic 10 (*Differential Item Functioning)* and Topic 13 (*Adaptive Testing and Test Assembly*). Then, when the interaction effects of year and journal type were analyzed, it was observed that there were no significant interaction effects for Topic 2 (*Validity and Equity in Educational and Psychological Testing*), Topic 6 (*Test Equating and Linking Methods*) and Topic 12 (*Rater Reliability and Generalizability*) and Topic 14 (*Mediation Effects and Multilevel Modeling)*. The graphs in Figure 6 illustrate the significant interaction effects between publication year and journal type. While not all journals are included for clarity, the analysis reveals several noteworthy trends. According to Figure 6, Topic 1 (*Parameter Estimation with Bayesian Methods)* has shown a significant increase over time in the EPM journal, although its overall prevalence remains low. Similarly, Topic 5 *(Item Response Theory)* has exhibited a rising trend in the EPM journal. In contrast, Topic 8 *(Scale Development and Validation),* despite being the most prevalent topic in the corpus, has experienced a notable decline in the EPM journal over the years.

Figure 6 also reveals that the JEBS journal has seen a pronounced decrease in the prevalence of Topic 3 *(Simulation-Based Power and Error Analysis).* While Topic 9 *(Formative Assessment and Feedback)* remains one of the most common topics, its prevalence has declined over time in the SEV journal, whereas it has remained stable in the EPM journal. On the other hand, Topic 10 (*Differential Item Functioning)*, the least prevalent topic in the corpus, has shown a slight increase over time in the EPM journal. Topic 11 *(Structural Equation Modeling),* one of the most prevalent topics, has the highest representation in the EPM journal, but its prevalence in the PSYCH journal has gradually declined. Finally, Topic 13 *(Adaptive Testing and Test Assembly)*, among the least prevalent topics, has demonstrated gradual growth in the EPOD and EPM journals.

Figure 6. *Expected topic proportions of year and journal type interactions effect.*



4. DISCUSSION and CONCLUSION

This study conducted an in-depth analysis of the psychometric literature by using STM to identify thematic trends, the evolution of these dynamics over time, and the differences in topics covered across journals. To accomplish this, eleven leading journals in the field were included in the analysis. As a result, 14 topics were identified and labeled by the researchers. The findings revealed that Topic 8 (Scale Development and Validation) emerged as the most dominant topic, while Topic 10 (Differential Item Functioning) had the lowest prevalence in the corpus. The prevalence values of the topics ranged from 5% to 10%, with four topics occupying the upper end of this spectrum. These results highlight the major areas of focus in psychometric research and illustrate shifts in interest over time.

Previous applications of topic modeling in educational and psychological domains (e.g., Anderson *et al.*, 2020; Wheeler *et al.*, 2024; Xiong & Li, 2023) have focused on specific contexts such as constructed-response items, validity evidence, or automated scoring systems. However, to the best of our knowledge, no prior study has systematically applied STM to a comprehensive set of psychometric publications over time and across multiple journals. Compared to existing literature, the present study contributes a broader, field-wide perspective by combining bibliometric reach with the explanatory capacity of STM. This approach enables researchers to trace not only the dominant topics but also their temporal dynamics and journal-level distributions, thus offering a deeper understanding of psychometric scholarship.

Topic 8 (Scale Development and Validation) and Topic 2 (Validity and Equity in Educational and Psychological Testing) stand out by playing a vital role in both the theoretical and practical dimensions of psychometric research. This suggests that scale development and validation continue to be central to psychometric research, particularly given their fundamental role in ensuring robust measurement instruments. However, the interaction effect analysis revealed that while Topic 8 remains dominant, its prevalence has decreased over time in the EPM journal. This trend may indicate that some journals are shifting their focus towards newer methodologies and applications. On the other hand, emerging topics such as Topic 2 (Validity and Equity in Educational and Psychological Testing), Topic 4 (Test-Taking Behavior and Cognitive Diagnosis Models), and Topic 13 (Adaptive Testing and Test Assembly) underscore the increasing focus on technology-driven methodologies and individualized assessment practices. This shift aligns with broader advancements in fields such as educational technology and machine learning, where adaptive and personalized approaches are becoming increasingly important (van der Linden & Glas, 2000; Borsboom, 2005). However, the current study identifies Topic 9 (Formative Assessment and Feedback) and Topic 14 (Mediation Effects and Multilevel Modeling) as additional 'hot' topics, showing increasing prevalence over the years. This indicates a growing emphasis on assessment processes that prioritize formative evaluation and advanced statistical modeling techniques.

Topic 2, Validity and Equity in Educational and Psychological Testing, highlights the evolving nature of validity discourse in response to social, ethical, and technological advancements. While traditional validity frameworks primarily focused on test revisions and validation processes (Plake & Wise, 2014; Cizek *et al.*, 2010), recent discussions emphasize the broader societal implications of assessment practices (Buzick *et al.*, 2023). The integration of racial justice perspectives into validity theory (Lederman, 2023; Randall *et al.*, 2022) underscores the necessity of ensuring fairness and equity in educational and psychological measurement. Additionally, the increasing reliance on automated scoring systems (Rupp, 2018) and artificial intelligence-driven assessments (Briggs, 2024) calls for renewed scrutiny regarding the transparency, ethical use, and bias mitigation in these technologies. As validity continues to expand beyond psychometric properties to encompass social responsibility, future research should explore the intersection of validity theory with emerging AI methodologies, ethical assessment practices, and the broader implications of test fairness in diverse populations.

Despite the declining interest in "cold" topics such as Topic 8 (Scale Development and Validation), Topic 12 (Rater Reliability and Generalizability), and Topic 11 (Structural Equation Modeling), it is essential to explore how these areas can be better integrated into contemporary practices. These fundamental domains remain critical for the validity and reliability of measurement tools. Research could focus on revisiting these issues and updating them using modern technologies, particularly in high stakes testing contexts. Interestingly, while previous research suggested a consistent decline in simulation studies (Topic 3, Simulation-Based Power and Error Analysis), the present study shows that simulation remains a critical tool, albeit with reduced prominence compared to real-data-driven approaches. The decline in simulation studies is particularly notable in the JEBS journal, which may indicate a preference for empirical data over simulated conditions in recent years.

Another key finding of this study is the network of correlations between topics. Topic 2 (Validity and Equity in Educational and Psychological Testing) was found to be strongly linked to Topic 7 (Exploring Predictors of Academic Performance) and Topic 9 (Formative Assessment and Feedback), suggesting an interrelationship between test validity, student performance, and formative assessment practices. Furthermore, Topic 1 (Parameter Estimation with Bayesian Methods) was connected to multiple topics, including Topic 3 (Simulation-Based Power and Error Analysis), Topic 4 (Test-Taking Behavior and Cognitive Diagnosis Models), Topic 5 (Item Response Theory), and Topic 11 (Structural Equation Modeling), reinforcing its foundational role in psychometric methodologies. Topic 1 can be regarded as a key concept that connects these topics due to its wide-ranging applications in psychometrics. Additionally, Topic 12 (Rater Reliability and Generalizability) and Topic 14 (Mediation Effects and Multilevel Modeling) were correlated, and both were connected to Topic 3 (Simulation-Based Power and Error Analysis). Furthermore, Topic 10 (Differential Item Functioning) shared connections with Topic 3 (Simulation-Based Power and Error Analysis) and Topic 5 (Item Response Theory), which is an expected result given that these methods can be employed together to enhance test fairness and reliability.

The potential of artificial intelligence and big data technologies to enhance psychometric modeling and the development of measurement tools should be explored. For instance, integrating AI-based algorithms to understand more complex data structures, model response patterns, and improve prediction accuracy in large-scale testing is crucial. The ethical implementation, transparency, and reliability of these technologies should also be prioritized in scholarly discourse. These recommendations have the potential to facilitate the advancement of innovative research and effective practices within the field of psychometrics. By capitalizing on the opportunities presented by emerging technologies while maintaining a strong adherence to the theoretical foundations of the discipline, psychometrics could evolve into a more equitable, efficient, and accessible field.

To enhance the effective integration of psychometric findings into educational policies and practices, it is essential to establish stronger connections between academic research and applied studies. Guidance on the application of emerging methodologies, such as cognitive diagnostic modeling, is critical for promoting equity, fairness, and accessibility, especially in the context of educational assessment systems. This study has illustrated the interconnections between specific topics, such as item response theory and test equating and linking methods. Future research could focus on interdisciplinary projects that further strengthen these connections. For example, future research could examine how various psychometric methods can be simultaneously integrated within test development processes. Additionally, this study is confined to abstracts of articles published in eleven journals indexed in the WoS database. The selected journals are recognized as prominent within the fields of measurement, evaluation, and psychometrics. Replicating this study with a broader range of journals may yield different results.

The STM analysis also revealed significant interaction effects between publication year and journal type. For instance, Topic 1 (Parameter Estimation with Bayesian Methods) and Topic 5 (Item Response Theory) have shown significant increases over time, particularly in the EPM journal. In contrast, Topic 11 (Structural Equation Modeling), despite its initial prominence, has exhibited a decline in the PSYCH journal. These findings suggest that the prevalence of specific psychometric topics may vary across journals, reflecting different editorial priorities and emerging research trends.

The widespread adoption of technologies such as big data analytics and machine learning has enhanced the accessibility and meaningfulness of real data analysis. Despite the decline in simulation studies, their critical role in strengthening theoretical and methodological foundations should not be overlooked. It is clear that simulation continues to be a critical tool, particularly for the testing and validation of new methodologies. In the future, establishing a balance between simulation studies and real data analyses may provide an approach that fosters both theoretical rigor and practical utility in psychometric research. In this regard, the growing prominence of real data could encourage the integration of simulation studies with more realistic scenarios and hybrid methods, thereby contributing to the development of more robust and comprehensive analytical frameworks within the field of psychometrics.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Kübra Atalay Kabasakal: Research Design, Resources, Methodology, Visualization, Software, Analysis, and Writing-original draft. **Duygu Koçak**: Research Design, Systematic review, Methodology, Supervision, and Critical Review. **Rabia Akcan:** Research Design, Systematic review, Methodology, Supervision, and Critical Review.

Orcid

Kübra Atalay Kabasakal https://orcid.org/0000-0002-3580-5568
Duygu Koçak https://orcid.org/0000-0003-3211-0426
Rabia Akcan https://orcid.org/0000-0003-3025-774X

REFERENCES

Ackerman, T.A., Bandalos, D.L., Briggs, D.C., Everson, H.T., Ho, A.D., Lottridge, S.M., Madison, M.J., Sinharay, S., Rodriguez, M.C., Russell, M., Von Davier, A.A., & Wind, S.A. (2023). Foundational competencies in educational measurement. *Educational Measurement Issues and Practice, 43*(3), 7–17. https://doi.org/10.1111/emip.12581

Anderson, D., Rowley, B., Stegenga, S., Irvin, P.S., & Rosenberg, J.M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice, 39*(4), 53-64. https://doi.org/10.1111/emip.12314

Bai, X., Zhang, X., Li, K.X., Zhou, Y., & Yuen, K.F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy, 102,* 11-24. https://doi.org/10.1016/j.tranpol.2020.12.013

Banks, G.C., Woznyj, H.M., Wesslen, R.S., & Ross, R.L. (2018). A review of best practice recommendations for text analysis in R (and a User-Friendly app). *Journal of Business and Psychology*, *33*(4), 445–459. https://doi.org/10.1007/s10869-017-9528-3

Bastola, M. N., & Hu, G. (2021). Chasing my supervisor all day long like a hungry child seeking her mother!: Students' perceptions of supervisory feedback. *Studies in Educational Evaluation*, *70*, 101055. https://doi.org/10.1016/j.stueduc.2021.101055

Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Non-normal data: Is ANOVA still a valid option? *Psicothema, 30*(4), 552-557. https://doi.org/10.7334/psicothema2018.245

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance, 6*(4), e21978.

Briggs, D.C. (2024). Strive for measurement, set new standards, and try not to be evil. *Journal of Educational and Behavioral Statistics*, *49*(5), 694-701. https://doi.org/10.3102/10769986241238479

Brooks, C., Burton, R., Van Der Kleij, F., Carroll, A., Olave, K., & Hattie, J. (2020). From fixing the work to improving the learner: An initial evaluation of a professional learning intervention using a new student-centred feedback model. *Studies in Educational Evaluation*, *68*, 100943. https://doi.org/10.1016/j.stueduc.2020.100943

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511613980

Buckhalt, J.A. (1999). Defending the science of mental ability and its central dogma. Review of Jensen on Intelligence-g-Factor. *Psycoloquy, 10*(23). http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?10.47

Buzick, H.M., Casabianca, J.M., & Gholson, M.L. (2023). Personalizing Large-Scale Assessment in practice. *Educational Measurement Issues and Practice*, *42*(2), 5–11. https://doi.org/10.1111/emip.12551

Chen, S., & Lei, P. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, *29*(3), 204-217. https://doi.org/10.1177/0146621604271495

Chen, J., Chen, C., & Shih, C. (2013). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18-36. https://doi.org/10.1177/0146621613488643

Choi, J.Y., Hwang, H., Yamamoto, M., Jung, K., & Woodward, T.S. (2016). A unified approach to functional principal component analysis and functional Multiple-Set canonical correlation. *Psychometrika*, *82*(2), 427–441. https://doi.org/10.1007/s11336-015-9478-5

Cizek, G.J., Bowen, D., & Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: a Follow-Up Study. *Educational and Psychological Measurement*, *70*(5), 732–743. https://doi.org/10.1177/0013164410379323

Cohn, S., & Huggins-Manley, A.C. (2019). Applying unidimensional models for semiordered data to scale data with neutral responses. *Educational and Psychological Measurement*, *80*(2), 242–261. https://doi.org/10.1177/0013164419861143

Jones, L.V., & Thissen, D.M. (2006). *A history and overview of psychometrics.* In Handbook of statistics (pp. 1–27). https://doi.org/10.1016/s0169-7161(06)26001-2

Gao, X., & Sazara, C. (2023). Discovering mental health research topics with topic modeling. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.13569

Göral, S., Özkan, S., Sercekus, P., & Alataş, E. (2021). The validity and reliability of the Turkish version of the Attitudes to Fer-Tility and Childbearing Scale (AFCS). *International Journal of Assessment Tools in Education, 8(4),* 764-774. https://doi.org/10.21449/ijate.773132

Gregson, T. (1991). The separate constructs of communication satisfaction and job satisfaction. *Educational and Psychological Measurement*, *51*(1), 39-48. https://doi.org/10.1177/0013164491511003

Groenen, P.J.F., & van der Ark, L.A. (2006). Visions of 70 years of psychometrics: the past, present, and future. *Statistica Neerlandica, 60*(2), 135–144. https://doi.org/10.1111/j.1467-9574.2006.00318.x

Guo, J., & Luh, W. (2008). Approximate sample size formulas for testing group mean differences when variances are unequal in One-Way ANOVA. *Educational and Psychological Measurement*, *68*(6), 959–971. https://doi.org/10.1177/0013164408318759

Hidalgo, M.D., & LÓPez-Pina, J.A. (2004). Differential Item Functioning Detection and Effect Size: A Comparison between Logistic Regression and Mantel-Haenszel Procedures. *Educational and Psychological Measurement*, *64*(6), 903-915. https://doi.org/10.1177/0013164403261769

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika*, *61*(1), 31-39. https://doi.org/10.1007/bf02296957

Hwang, S., Flavin, E., & Lee, J.E. (2023). Exploring research trends of technology use in mathematics education: A scoping review using topic modeling. *Education and Information Technologies, 28*, 10753–10780. https://doi.org/10.1007/s10639-023-11603-0

Jiang, Y., Von Davier, A.A., & Chen, H. (2012). Evaluating equating results: percent relative error for chained kernel equating. *Journal of Educational Measurement*, *49*(1), 39–58. https://doi.org/10.1111/j.1745-3984.2011.00159.x

Jiang, X., & Ironsi, S.S. (2024). Do learners learn from corrective peer feedback? Insights from students. *Studies in Educational Evaluation*, *83*, 101385. https://doi.org/10.1016/j.stueduc.2024.101385

Kiers, H.A.L. (1997). Three-mode orthomax rotation. *Psychometrika*, *62*(4), 579–598. https://doi.org/10.1007/bf02294644

Kim, S. (2001). An evaluation of a Markov Chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, *25*(2), 163-176. https://doi.org/10.1177/01466210122031984

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education, 36*(3), 242-254. https://doi.org/10.1080/08957347.2023.2214654

Liu, J., & Low, A.C. (2008). A Comparison of the Kernel Equating Method with Traditional Equating Methods Using SAT® Data. *Journal of Educational Measurement*, *45*(4), 309–323. https://doi.org/10.1111/j.1745-3984.2008.00067.x

MacDonald, P.L., & Gardner, R.C. (2000). Type I Error Rate Comparisons of Post Hoc Procedures for I j Chi-Square Tables. *Educational and Psychological Measurement*, *60*(5), 735–754. https://doi.org/10.1177/00131640021970871

Martin, C.R., & Savage-McGlynn, E. (2013). A 'good practice' guide for the reporting of design and analysis for psychometric evaluation. *Journal of Reproductive and Infant Psychology, 31*(5), 449–455. https://doi.org/10.1080/02646838.2013.835036

Meeter, M. (2022). Predicting Retention in Higher Education from high-stakes Exams or School GPA. *Educational Assessment*, *28*(1), 1-10. https://doi.org/10.1080/10627197.2022.2130748

Michell, J. (2022). The art of imposing measurement upon the mind: Sir Francis Galton and the genesis of the psychometric paradigm. *Theory & Psychology*, *32*(3), 375-400. https://doi.org/10.1177/09593543211017671

Pan, Y., Livne, O., Wollack, J.A., & Sinharay, S. (2023). Item selection algorithm based on collaborative filtering for item exposure control. *Educational Measurement Issues and Practice*, *42*(4), 6–18. https://doi.org/10.1111/emip.12578

Park, S., Steiner, P.M., & Kaplan, D. (2018). Identification and sensitivity analysis for average causal mediation effects with time-varying treatments and mediators: Investigating the underlying mechanisms of kindergarten retention policy. *Psychometrika*, *83*(2), 298–320. https://doi.org/10.1007/s11336-018-9606-0

Plake, B.S., & Wise, L.L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement Issues and Practice, 33*(4), 4–12. https://doi.org/10.1111/emip.12045

Polatgil, M. (2023). Analyzing comments made to the Duolingo mobile application with topic modeling. *International Journal of Computing and Digital Systems, 13*(1), 223–230.

Randall, J., Slomp, D., Poe, M., & Oliveri, M.E. (2022). Disrupting White Supremacy in Assessment: Toward a Justice-Oriented, Antiracist validity framework. *Educational Assessment, 27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Richardson, G.M., Bowers, J., Woodill, A.J., Barr, J.R., Gawron, J.M., & Levine, R.A. (2014). Topic models: A tutorial with R. *International Journal of Semantic Computing, 8(01), 85-98.*

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., & Rand, D.G. (2014). Structural topic models for Open-Ended Survey Responses. *American Journal of Political Science, 58*(4), 1064-1082. https://doi.org/10.1111/ajps.12103

Roberts, M.E., Stewart, B.M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Rupp, A.A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, *31*(3), 191–214. https://doi.org/10.1080/08957347.2018.1464448

Schuster, C. (2004). A Note on the Interpretation of Weighted Kappa and its Relations to Other Rater Agreement Statistics for Metric Scales. *Educational and Psychological Measurement*, *64*(2), 243–253. https://doi.org/10.1177/0013164403260197

Shih, C., & Wang, W. (2009). Differential Item Functioning Detection Using the Multiple Indicators, Multiple Causes Method with a Pure Short Anchor. *Applied Psychological Measurement*, *33*(3), 184–199. https://doi.org/10.1177/0146621608321758

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open-Source Software*, *1*(3), 37. https://doi.org/10.21105/joss.00037

Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, *48*(2), 157–217. https://doi.org/10.1007/s10462-016-9498-2

Sireci, S.G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99–104. https://doi.org/10.1111/jedm.12005

Sijtsma, K., & Pfadt, J.M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika, 86*(4), 843-860.

Tharenou, P., & Terry, D.J. (1998). Reliability and validity of scores on scales to measure managerial aspirations. *Educational and Psychological Measurement*, *58*(3), 475–492. https://doi.org/10.1177/0013164498058003008

Talloen, W., Moerkerke, B., Loeys, T., De Naeghel, J., Van Keer, H., & Vansteelandt, S. (2016). Estimation of indirect effects in the presence of unmeasured confounding for the Mediator–Outcome relationship in a multilevel 2-1-1 mediation model. *Journal of Educational and Behavioral Statistics*, *41*(4), 359-391. https://doi.org/10.3102/1076998616636855

Tonidandel, S., Summerville, K.M., Gentry, W.A., & Young, S.F. (2021). Using structural topic modeling to gain insight into challenges faced by leaders. *The Leadership Quarterly*, *33*(5), 101576. https://doi.org/10.1016/j.leaqua.2021.101576

Tunç, E.B., Parlak, S., Uluman, M., & Eryiğit, D. (2021). Development of the Hostility in Pandemic Scale (HPS): A Validity and Reliability study. *International Journal of Assessment Tools in Education*, *8*(3), 475–486. https://doi.org/10.21449/ijate.837616

Wheeler, J.M., Cohen, A.S., & Wang, S. (2024). A comparison of latent semantic analysis and latent Dirichlet allocation in educational measurement. *Journal of Educational and Behavioral Statistics, 49*(5), 848–874. https://doi.org/10.3102/10769986231209446

Van Der Ark, L.A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, *70*(2), 283–304. https://doi.org/10.1007/s11336-000-0862-3

Van der Linden, W.J., & Glas, C.A.W. (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/978-1-4757-3224-0

Vitoratou, S., & Pickles, A. (2017). Psychometric analysis of the Mental Health Continuum-Short Form. *Journal of Clinical Psychology, 73*(10), 1307-1322. https://doi.org/10.1002/jclp.22422

Xiong, J., & Li, F. (2023). Bilevel topic model-based multitask learning for constructed-response multidimensional automated scoring and interpretation. *Educational Measurement: Issues and Practice, 42*(2), 42–61. https://doi.org/10.1111/emip.12550

Yavuz, S., Odabaş, M., & Özdemir, A. (2016). Öğrencilerin sosyoekonomik düzeylerinin TEOG matematik başarısına etkisi [Effect of socio-economic status on student's TEOG mathematics achievement]. *Journal of Measurement and Evaluation in Education and Psychology, 7*(1), 85–95. https://doi.org/10.21031/epod.86531

Zagaria, A., & Lombardi, L. (2024). Bayesian versus frequentist approaches in psychometrics: a bibliometric analysis. *Discover Psychology, 4, 61*. https://doi.org/10.1007/s44202-024-00164-z

Zhan, P., Man, K., Wind, S.A., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, *47*(6), 736–776. https://doi.org/10.3102/10769986221111085