Modeling demands-resources fit in teacher education using open-ended data: a methodological-substantive synergy

Fernando Núñez-Regueiro, Samuel Falcon, Pascal Bressoux

Education and Information Technologies
https://doi.org/10.1007/s10639-025-13764-6

Modeling demands-resources fit in teacher education using open-ended data: a methodological-substantive synergy

Fernando Núñez-Regueiro[1] · **Samuel Falcon**[2,3] · **Pascal Bressoux**[1]

Received: 1 September 2024 / Accepted: 19 August 2025

Abstract

This study explores the effectiveness of large language models (LLMs) in automatically encoding a large set of open-ended responses to obtain data for use in applied statistics. As a case study, we focus on demands-resources fit processes and engagement in teacher education. To probe the validity of LLMs in investigating these processes, we compare results from measures obtained via ordinary Likert-type items (scale measures), and measures obtained from automatically encoding open-ended questions (LLM measures) for the same sample of student teachers (N=499, 82% female, M_{age}=23.5 years). Results demonstrate the reliability of LLMs in processing and quantifying large amounts of open-ended data quickly and as accurately as scale measures. Moreover, results concur to reveal an "optimal margin" of demands-resources fit in student teacher engagement. Accordingly, study resources surpassing study demands maximizes engagement, whereas insufficient resources minimize it, and moderate levels of both demands and resources lead to intermediate engagement. By contrast, high or low levels of both demands and resources are suboptimal for engagement. Taken together, these findings demonstrate that LLM-derived statistics offer an efficient and reliable approach to extracting data from open-ended responses, enabling the large-scale analysis of qualitative insights while preserving their richness. This method facilitates the integration of qualitative and quantitative approaches, enhancing the study of individual behavior, and holds significant potential for enhancing digital education frameworks by supporting adaptive learning systems and digital assessment practices.

Keywords Study demands-resources fit · Large language models · Cubic response surface analysis · AI in education · Data analysis automation

Extended author information available on the last page of the article

Published online: 06 October 2025

1 Introduction

Educational studies, including research into teaching practice or teacher education, are based on two major kinds of data collection methods, that is, structured and open-ended methods. Structured methods relate to observables or, more often, self-reported items answered via closed-ended response scales. Because the response scale is the same for all respondents, the scores derived from the scales can be used as quantitative indicators of individual differences about the phenomenon being measured. Such indicators allow for quantitative models to emerge, but they come at the cost of ignoring information about other elements of the phenomenon not accounted for by the closed-ended response scales (e.g., facets not tapped by the included items). Conversely, open-ended collection methods use interviews or questions with an open-ended format to retrieve information about the full scope of experiences recalled by respondents. Open-ended information occupies a central place in the field of research, providing deep and contextual insights that are often not available through structured data collection methods alone (Kandel, 2020). For instance, in studies of human behavior and social phenomena, open-ended information can uncover nuances and complexities that quantitative approaches might miss (Creswell & Creswell, 2017). However, open-ended data collection methods bring with them the challenge of an extensive manual encoding process to transform the information into analyzable data (Mohajan, 2018). The time and cost associated with manual encoding often prevents researchers from scaling up their samples and performing advanced statistical analyses (Lennon et al., 2021). Therefore, historically, open-ended data has rarely been used as a basis for statistical modeling, more often confined to the domain of qualitative data analyses.

Yet, recent developments have revealed new research perspectives based on open-ended data collection methods. The development and adoption of new technologies, such as the use of artificial intelligence (AI), has enabled to automate and streamline the process of encoding this type of data (Kaufmann et al., 2020; Leeson et al., 2019; Rietz & Maedche, 2021). By applying AI-based models, researchers can analyze large volumes of textual data efficiently while maintaining the richness and depth that characterizes open-ended data (Chang et al., 2021). This innovation also allows for an integration of open-ended and structured data collection approaches, opening up new possibilities in the field of social research (Demszky et al., 2023). However, despite these advances, there is a notable gap in the existing literature concerning the use of these techniques for more than just testing the reliability of the method. To our knowledge, no study has combined these techniques with more conventional self-reported scales to measure specific variables of interest. This integration is crucial for verifying the validity of data derived from automatically encoded open-ended information and facilitating diverse data analyses on large sample sizes.

Building on this, the present study aims to advance both methodology and substantive research by leveraging Large Language Models (LLMs) to automatically encode open-ended responses and applying this approach to investigate demands-resources fit in teacher education (Salmela-Aro et al., 2022). In this way, rather than treating LLM-based text encoding as an isolated methodological development, we adopt a methodological-substantive synergy (Marsh, 2007; Marsh & Hau, 2007), where the

refinement of an analytical technique is directly tied to enhancing our understanding of a substantive issue. This synergy has previously been illustrated in many studies, such as, for example, the one conducted by Marsh et al. (2009), in which they combined multilevel modeling and latent-variable modeling to analyze the Big-Fish-Little-Pond Effect, highlighting contextual influences on academic self-concept. In our study, methodologically, we assess the validity of LLM-derived measures by comparing them with traditional self-reported scales of study demands and resources. Substantively, we use these measures to examine how the interplay between study demands and resources shapes student teachers' engagement in teacher education. We chose this framework because it has been widely used in previous research, and its characteristic nonlinear patterns, such as the optimal margin of fit (Núñez-Regueiro et al., 2022, 2025; Dupéré et al., 2015), provide a useful reference point for evaluating whether similar results can be obtained using LLM-based data. Therefore, by integrating structured and open-ended data sources, this study not only addresses unresolved questions about the predictive validity of automatic encoding with LLMs but also contributes to the broader theoretical discourse on demands-resources fit, offering a better understanding of student teacher engagement processes.

1.1 Understanding behavior using open-ended data: strengths and limitations of current approaches

Analysis of open-ended information is essential for comprehending human behavior (Johnson & Onwuegbuzie, 2007). This form of analysis, as Onwuegbuzie et al. (2012) highlight, offers a nuanced understanding of experiences by delving into various types of data, including discourse studies, observations, and diverse media such as drawings and videos. Written documents, encompassing text productions, audio transcripts, and responses to open-ended questions, are notably prevalent in research. This trend is substantiated by the extensive development of text analysis tools (Kuckartz, 2014), reflecting the scientific community's commitment to extracting meaningful insights from such data (Núñez-Regueiro & Núñez-Regueiro, 2021; Falcon & Leon, 2023; Abdullah et al., 2023; Usai et al., 2018).

The strength of text analysis lies in its diverse methodologies. Traditional approaches include semiotics, which interpret signs and symbols in communication; comparative analysis, identifying patterns across case studies; and keywords-in-context, focusing on the usage and context of specific words within texts. These methods, alongside classical content analysis, provide a comprehensive view of the data (Leech & Onwuegbuzie, 2008). By employing these varied techniques, researchers can conduct in-depth thematic analysis, fostering abductive reasoning that facilitates the development of new theories, or the modification of existing ones based on empirical evidence (Walker & Myrick, 2006). Text analysis also allows the capture of spontaneous, non-preformatted responses, offering a richness in data that structured scales or fixed-response options may overlook (Bailey, 2008). Therefore, text analysis, through its multifaceted approaches, plays a vital role in enriching our understanding of complex human experiences.

However, while these traditional methods offer profound insights into behavior, they are primarily limited by the labor-intensive nature of manual encoding (Rah-

man, 2016). Encoding large datasets manually requires multiple encoders and takes considerable time, which results in relatively small sample sizes (Marshall et al., 2013). Such limitations in sample size restrict the scope of data analysis, confining it to simpler forms, like descriptive analysis, thereby limiting the depth and breadth of research findings (Kim et al., 2017; Lennon et al., 2021). This scenario underscores the need for a research approach that retains the strengths of open-ended collection methods, such as the ability to capture rich and spontaneous responses, while also facilitating the analysis of larger sample sizes suitable for more advanced statistical analysis. By striking this balance, research can not only maintain the depth of open-ended data but also expand its potential through comprehensive statistical evaluation to generalize the findings (Onwuegbuzie & Daniel, 2003).

1.2 Methodological focus: taking open-ended data analysis to the next level using automatic data encoding

Recent advancements in natural language processing are transforming text analysis, particularly with the emergence of pre-trained Large Language Models (LLMs). These models leverage deep learning techniques to handle complex tasks that were once infeasible (Chang et al., 2021; Minaee et al., 2021). The introduction of LLMs like Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) or Generative Pretrained Transformer (GPT; Brown et al., 2020) has revolutionized the field of text analysis in social sciences, enabling applications in emotion recognition (Acheampong et al., 2021), generation of scale items (Götz et al., 2023), and identification of suicidal ideation in social media posts (Ophir et al., 2020), among many others. Beyond these applications, one of the most promising frontiers in behavioral research is automatic text encoding (Alqassab & Leon, 2024; Chiang & Lee, 2023; Jackson et al., 2022; Santana-Monagas et al., 2024).

LLMs are particularly competent at classifying texts without the need to train a model on human-labelled data, avoiding the traditional problems of creating costly training datasets (Brown et al., 2020; Mu et al., 2012). These models can accurately classify text based on clear, predefined instructions, making them an efficient alternative to manual coding (Álvarez-Álvarez & Falcon, 2023; Kasneci et al., 2023). A particularly effective technique in this domain is the "few-shot learning", where LLMs improve their performance by incorporating a few example entries into the initial instructions. For instance, text classification accuracy improves significantly by adding some classification examples to the base instruction (Brown et al., 2020). Given these results, social science research has increasingly leveraged this technique for text classification, demonstrating their ability to handle complex, theory-driven classifications with high precision (Demszky et al., 2023b; Kasneci et al., 2023; Le Mens et al., 2023).

Recent studies confirm that LLMs reliably classify open-ended responses, achieving results similar to humans with significantly fewer resources (Alqassab & Leon, 2024; Mellon et al., 2023; Santana-Monagas et al., 2024). For instance, Álvarez-Álvarez and Falcon (2023) utilized Smith and Baik's (2021) teaching practice definitions to code student responses about preferred teaching methods with an LLM. Their analysis revealed that the LLM's accuracy in categorizing responses aligned

with the levels of human agreement; higher human agreement concurred with more precise LLM classification, while lower agreement coincided with decreased accuracy, suggesting that the discrepancies were due to the category definitions rather than the LLM's capabilities. Based on the above evidence, this methodology promises numerous advantages over traditional manual encoding. Firstly, it offers reliable results, showing error rates similar to that of human encoding. Secondly, it is cost-effective, enabling the analysis of large samples that were previously impractical due to resource constraints (Rathje et al., 2023). This, in turn, may allow researchers to handle substantial volumes of data for the accumulation of samples large enough to support advanced statistical analyses. These benefits underscore the potential of automatic encoding using LLMs in transforming open-ended research.

However, recent critiques have argued that assessing the reliability of LLMs solely through agreement with human coders may be insufficient, particularly given their potential input fluctuations and occasional inaccuracies (Raj et al., 2023; Yu et al., 2023). As a result, it is increasingly recommended to broaden the scope of validation methods. This includes not only verifying interrater reliability but also evaluating the psychometric quality of LLM-derived data using complementary approaches such as convergent validity and concurrent validity (see Method; Núñez-Regueiro et al., 2021; Raykov & Marcoulides, 2011). For example, comparing data from LLM-encoded open-ended questions with data obtained through traditional scales can help assess whether both sources reflect similar constructs and relationships (Glazier et al., 2021). Moreover, integrating LLM-based and scale-based data into statistical models raises specific challenges, particularly in ensuring their interpretability and robustness in real-world research contexts (Östlund et al., 2011; Schilke et al., 2012). These considerations call for the use of inferential analyses, such as comparisons of structural patterns or predictive validity, when evaluating the overall usefulness and applicability of LLM-derived measures (Bacher-Hicks et al., 2019; Campbell et al., 2020).

Therefore, the literature reveals significant gaps concerning the reliability of LLMs in encoding open-ended questions, the comparative validity of such data against scale-derived data, and its usefulness for statistical analysis. Our study aims to address these gaps within the demands-resources model (D-R), a relevant theoretical framework in educational research.

1.3 Substantive focus: how study demands-resources fit relates to engagement in teacher education

The study demands-resources model (D-R) of student engagement and burnout (Salmela-Aro et al., 2022; Salmela-Aro & Upadyaya, 2014) provides a pertinent framework with real-world applicability. This model conceptualizes the positive and negative influences, respectively, of study demands (or "sources of stress") and study resources (or "sources of motivation") on student engagement. In the context of teacher education, study demands relate to academic factors that undermine student engagement, such as the workload arising from cumulative tasks (e.g., passing college examinations, writing a dissertation, preparing a teacher recruitment contest) or from the perception that classes are too numerous or not relevant for teaching

(Núñez-Regueiro & Leroy, 2024; Núñez-Regueiro et al., 2024; Clarke et al., 2012; Deasy et al., 2016; Flores & Niklasson, 2014). By contrast, study resources that contribute to student engagement in teacher education include self-efficacy to succeed academically (Núñez-Regueiro & Leroy, 2024; Löfström et al., 2010), feeling supported by teacher educators (Núñez-Regueiro & Leroy, 2024; Núñez-Regueiro et al, 2024; E. Kim & Corcoran, 2018), but also reporting an intrinsic motivation to become a teacher (Núñez-Regueiro & Leroy, 2024; Watt & Richardson, 2007).

While prior research has focused on the additive (linear) effects of study demands and resources on student engagement, a focus in line with original conceptualizations of D-R fit that underlined the nonexistence of interactive effects (Demerouti et al., 2001), more recent theoretical advancements have proposed to enlarge the scope of analysis, by investigating their synergistic or interactive effects (Bakker & Demerouti, 2017; Salmela-Aro et al., 2022). These advancements can build on individual stress theories that have long articulated how the interplay between contextual demands and individual resources affects motivation and engagement in the organizational and educational contexts (Núñez-Regueiro, 2017; Núñez-Regueiro et al., 2022, 2025; Dupéré et al., 2015; Edwards, 1996; Lazarus, 1999; LeCompte & Dworkin, 1991; Pearlin, 1999). In this circumplex approach, individual behavior is appraised as depending on the combination of levels of resources and demands (see Fig. 1A; Núñez-Regueiro & Wang, 2025). These levels may run in opposite directions and create D-R misfit, either positively in the form of energy ($resources = -demands > 0$; upper-left quadrant, Fig. 1A), or negatively in the form of exhaustion ($resources = -demands < 0$; lower-right quadrant, Fig. 1A). On the contrary, these levels may run in the same direction and create D-R fit. When the fit is positive ($demands = resources > 0$, upper-right quadrant, Fig. 1A), high demands are met with the activation of high resources, thus resulting in a state of tension. When the fit is negative ($demands = resources < 0$, lower-left quadrant, Fig. 1A), the activation of resources is low and reflects a state of calmness.

Until recently, D-R fit processes primarily involved moderation analysis, which modeled the interaction between demands (x) and resources (y) in terms of their additive and interaction effects on student engagement (see solid arrows, Fig. 1B; Núñez-Regueiro, 2018; Núñez-Regueiro et al. 2022; Yang et al., 2018). These studies revealed that some resources, like teacher support, academic self-efficacy, and a positive school climate, could mitigate the deleterious impact of school demands or stressors on student engagement at school or university (e.g., study workload, unclear school assignments, unfavorable school tracking, school bullying). While informative, these moderation analyses did not reveal whether this mitigation was observed for situations of misfit (e.g., higher resources than demands), or situation of fit (e.g., high resources mitigating high demands), or whether the mitigation depended on the degree of fit or misfit (e.g., nonlinearities associated with optimal or suboptimal margins of fit or misfit). This valuable information can be obtained by conducting a response surface analysis based on cubic polynomial regression (Núñez-Regueiro & Juhel, 2022, 2024; Humberg et al., 2020), which includes higher-order effects on student engagement (e.g., quadratic, cubic, interactive-quadratic effects; see dashed arrows, Fig. 1B). For example, preliminary findings using independent datasets and alternative scales in school and university contexts (Núñez-Regueiro et al., 2025)

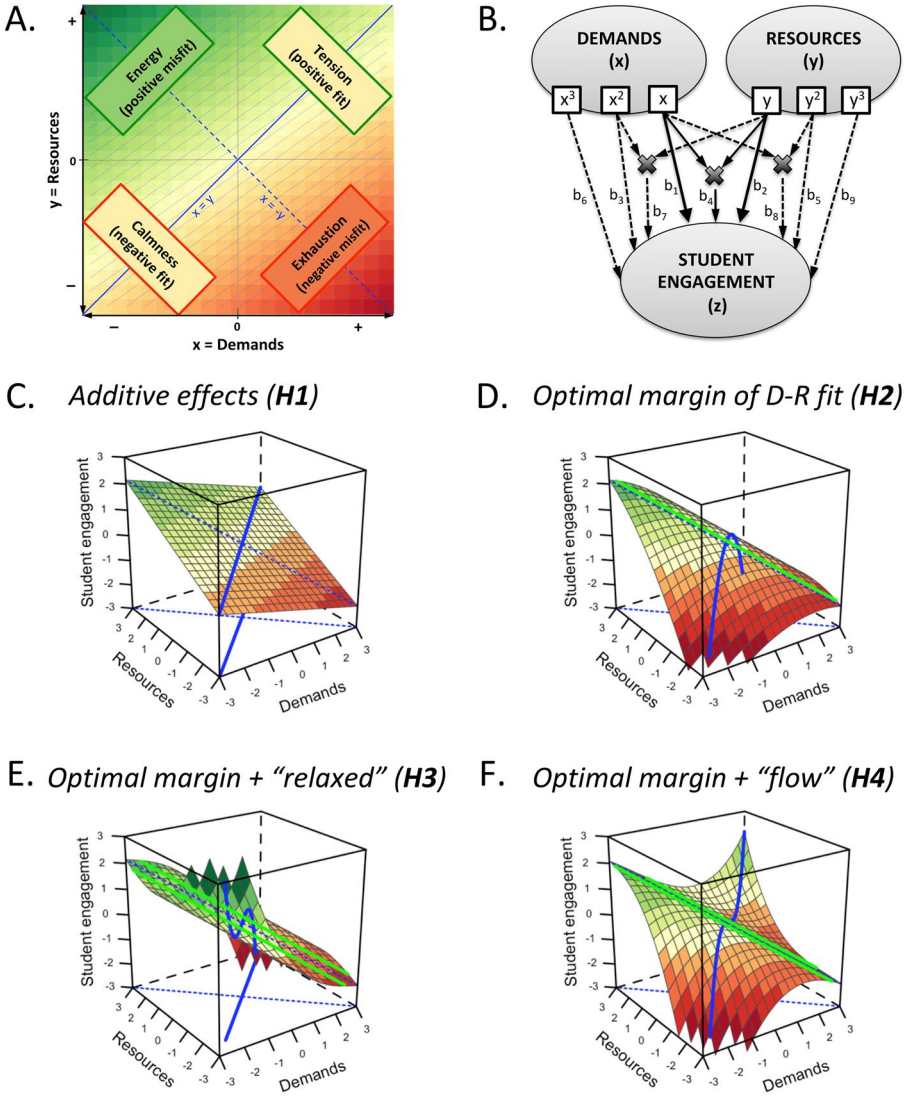

Fig. 1 Circumplex (**A**), Cubic Polynomial Regression (**B**), and Hypothetical Responses Surfaces (**C**-**F**) of Demands-Resources Fit Processes Undergirding Student Engagement. Note. The interplay between student demands and resources is conceived in the form of a bidimensional demands-resources circumplex (**A**), and related to student engagement using cubic polynomial regression modeling (**B**). Using this modeling approach, alternative response surfaces are hypothesized about linear effects (**C**) or interactive-nonlinear effects (**D**, **E**, **F**) of demands and resources on student engagement. Lines in green represent margins where the response behavior reverses in valence

revealed the existence of 3 paradigmatic D-R fit processes (see Fig. 1D and F; see also Choi, 2004). These findings suggest that having more resources than demands (positive misfit) generally relates positively to student engagement. Conversely, balanced resources and demands exhibit characteristic nonlinearities; the most typi-

cal being a quadratic relationship where moderate levels of D-R fit are optimal for engagement, while high or low levels induce either boredom or burnout (optimal margin hypothesis; Fig. 1D). Two variants of this pattern were also observed in which the latter extreme states were found to be more adaptive, either by inducing a state of "relaxation" in negative fit (optimal margin and "relaxed" hypothesis; Fig. 1E), or a state of "flow" in positive fit (optimal margin and "flow" hypothesis; Fig. 1F). Although these variants showed a better fit to the data, they were reasonably well approximated by the optimal margin pattern (Fig. 1D; Núñez-Regueiro et al., 2025; Choi, 2004).

Overall, the characteristic nonlinearities of D-R fit processes found via traditional scale measures (see Fig. 1D and F) offer a meaningful reference to evaluate whether similar patterns emerge from LLM-derived data. Identifying such surfaces would support the idea that automatic encoding preserves key theoretical properties of the constructs under study.

1.4 Present study

The present study is a substantive-methodological synergy aiming to model demands-resources fit processes of student teachers' engagement, based on open-ended data automatically encoded into count data by means of LLMs. Methodologically, the goal is to evaluate the quality of LLM-derived data, by conducting a reliability analysis of data encoded by the LLM and human encoders, and by assessing the convergent and concurrent validity of LLM measures as compared to scale measures of demands and resources. Substantively, this study aims to shed light on the relation between student teachers' demands and resources and their engagement in teacher education, particularly as it relates to the interplay between demands and resources in the form of fit or misfit processes using cubic RSA (see Fig. 1). Two research questions will guide the research:

RQ1: How reliable and valid is LLM-based encoding for quantifying open-ended responses in educational research?
RQ2: What insights do LLM-processed demands-resources data reveal about student teachers' engagement?

Regarding RQ1, LLM measures will be considered reliable if the agreement between the coding performed by the LLM and by human coders of a subset of data is satisfactory. In addition, according to criteria used to evaluate scales (DeVellis, 2016; Raykov & Marcoulides, 2011), LLMs will be deemed valid if the resulting measures are reliable, correlate systematically with alternative measures of demands and resources (convergent validity with scale measures), and bestow similar structural relations to other constructs as these alternative measures (concurrent validity).

Concerning RQ2, according to standard D-R fit theory (Demerouti et al., 2001; Salmela-Aro & Upadyaya, 2014), demands and resources were expected to have (linear) negative and positive effects on student teachers' engagement, respectively (H1; Fig. 1C). Alternatively, according to recent advancements in D-R fit theory (Núñez-Regueiro et al., 2025; Salmela-Aro et al., 2022), student teachers' engagement was

expected to respond to an optimal margin of D-R fit, with possible "relaxation" or "flow" effects of D-R fit (H2 to H4; Fig. 1D to F).

2 Methods

2.1 Participants and procedure

We collected data from a longitudinal sample of 579 student teachers (81.9% female, mean age = 23.4 years) from 22 training centers in France. Participants were first year students from a two-year national teacher education program, 58% of them preparing to teach in primary schools (42% in secondary schools). After approval by training centers, participants were invited to take a series of online questionnaires measuring demands and resources experienced in teacher education, and multiple indicators of motivation and engagement in teacher education (see Measures). Although participation rates were high across first-year measurements (T1 = 73%, T2 = 71%, T3 = 56%), the responses to open-ended questions were more limited (T1 = 31%, T2 = 31%, T3 = 22%). To increase coverage of the sample, it was decided to merge responses from the 3 measurements, by augmenting T1 respondents with (non-redundant) new respondents from T2 and T3. Thus, the final sample comprised 499 student teachers (82% female, mean age = 23.5 years), 62% of whom had answered one of the open-ended questions for LLM measures (see 2.3.3. Missing Data Treatment).

2.2 Measures

2.2.1 Predictors based on open-ended data: LLM measures

To obtain LMM-derived measures of academic demands and resources reported in writing by student teachers, we followed a methodology that utilized AI for analyzing responses to open-ended questions (Hynninen et al., 2019). First, we used open-ended data from a similar study to refine the prompt for the LMM and assess its reliability for encoding counts. This preliminary data used the same open-ended questions and focused on the same category of students as the data used for subsequent analyses (i.e., first-year teacher education students).

For coding these responses, we utilized the GPT-3.5 model, specifically the gpt-3.5-turbo version, setting the temperature to 0 and Top P to 1 for deterministic, replicable outputs (Álvarez-Álvarez & Falcon, 2023; Brown et al., 2020; OpenAI, 2024). Temperature controls randomness, with 0 ensuring consistent outputs, while Top P limits word selection diversity; setting it to 1 disables sampling, maximizing predictability.

The final prompt structure comprised four parts: (1) an initial instruction for classification, (2) definitions of categories comprising five representative examples of each category (i.e., academic resources or demands), (3) response format instructions (aiming for an Excel-like column output and count data), and (4) the student's response to be analyzed.

Second, having established its reliability (see Results), the above prompt was used to encode the open-ended questions for the present study. The instructions provided to GPT for classifying the answers on study demands and study resources remained identical to those for the preliminary data, except for the second part, which changed between them. This variation was necessary because the second part of the prompt included the definitions of categories. Therefore, when analyzing responses to the question about demands, we applied the definitions specific to demands; similarly, we used the definitions specific to resources when examining the resource-related answers.

Moreover, we automated the encoding using a Python script and three input files, namely two text files containing the first three parts of the prompt and essential model details (like the API key and hyperparameters), and a CSV file with student responses (fourth part of the prompt; see Fig. 2). The script's tasks included assembling complete prompts for each student's response, sending these prompts to OpenAI's servers, and then collecting and storing the model's responses in a new CSV file. This output file mirrored the original response file but with an additional column indicating the classification of each student's answer. This approach enabled counting occurrences of academic demands and resources reported in writing by each student, using the following open-ended questions and definitions of categories:

Study demands (LLM) Demands were described in writing by the student teachers in response to the open-ended question: "*If you have any comments about sources of stress or difficulties in your training*, *please let us know below*". Those demands that were academic in nature were then identified by LMMs using the definition: "*Intrinsic and extrinsic elements of the academic process*, *such as training*, *workload*, *assessments*, *interactions with supervisors and peers*, *and demands of the educational system*".

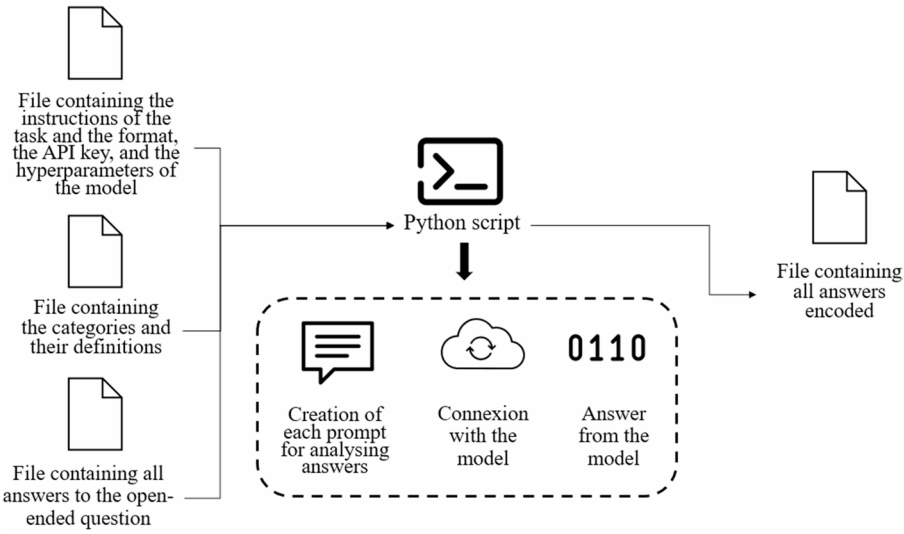

Fig. 2 Python script for answer analysis

Study resources (LLM) Similarly, students reported their resources by answering the open-ended question: "*If you have any comments about sources of motivation in your training*, *please let us know below*". We then identified responses to this question as occurrences of academic motivations when they aligned with the following definition: "*Factors stimulating the desire to learn*, *including elements related to course structure*, *support from supervisors*, *peer success*, *subject interest*, *personal success goals*, *and student mutual support*".

2.2.2 Predictors based on structured data: scale measures

Scale-derived measures of academic demands and resources were obtained using multiple items answered on a 6-point Likert-type response scale (1 = "Do not agree at all" to 6 = "Totally agree"). These closed-ended data provided evidence on study demands and study resources processes in teacher education, as modeled by standard quantitative approaches. This evidence provided a benchmark against which the evidence provided by models based on LLM-measures was evaluated.

Study demands (Scales) Scale measures of study demands were obtained by asking students to report their perceived study workload, based on 4 items tapping the amount of study hours [*The personal workload (reading books and articles*, *apprenticeships*, *etc.) is very heavy]*, the accumulation of tasks (*There are lots of assignments to hand in or exams to take*, *The accumulation of tasks in the master's program (assignments*, *mid-term exams*, *competitions*, *internships*, *etc.) is hard work*), and the amount of classes (*There are many hours of lectures or tutorials)*. The scale showed satisfactory internal consistency ($\alpha = .767$).

Study resources (Scales) Study resources were measured in terms of perceived support from teacher educators (stem question: *which amount of the following do you experience in your training?*), using 4 items underscoring support for needs for autonomy (*Trainers who try to understand how students see things before suggesting a new way of doing things*, *Trainers who take into consideration the students' way of doing things*), competence (*Trainers who ensure that students really understand the objectives of the work and what needs to be done*), and relatedness (*The feeling of being understood by trainers*). We focused on teacher support based on the observation that student teachers have reported this factor as a major resource in their teacher education training (Núñez-Regueiro et al., 2024), and because this factor is more impactful than other academic sources of social support (e.g., peer support) in processes of student engagement (Núñez-Regueiro & Leroy, 2024; Núñez-Regueiro et al., 2025; Furrer & Skinner, 2003). The internal consistency of the teacher support scale was excellent ($\alpha = .905$).

2.2.3 Outcome

Student teachers' engagement Aligning with conceptualizations of student engagement as a metaconstruct denoting commitment to teacher education and to the teaching career (Kahu, 2013; Roberts, 2012), student teachers' engagement was measured

based on their emotional engagement ($\alpha = .820$, 4 items rated on a 7-point scale, sample item=*I am enthusiastic about my training*; adapted from Zecca et al., 2015), satisfaction with career choice ($\alpha = .801$, 4 items rated on a 6-point Likert scale, sample item=*Generally speaking, I am satisfied with my decision to become a teacher*; adapted from OECD, 2019; Watt & Richardson, 2007), and study career turnover intentions ($\alpha = .870$, 3 items rated on a 7-point scale, reversed-coded, sample item=*Do you ever seriously think about changing your career plans?*; adapted from O'Reilly et al., 1991). These factors loaded on a second-order factor of student engagement in teacher education ($0.611 < ß < 0.911$) in a way that provided excellent fit to the data [$\chi^2(41) = 85.6$, $p < .001$, CFI = 0.982, RMSEA = 0.051, SRMR = 0.039].

2.2.4 Covariates for network analysis

Covariates The questionnaire also contained measures for covariates that were used as part of the covariance structure for the network analyses. These measures included student teachers' perceived peer support ($\alpha = .891$, 4 items, sample item=*Developing close ties with other students from the program*; adapted from Johnson & Norem-Hebeisen, 1979), value of the training ($\alpha = .751$, 3 items, sample item=*The courses of my training will come in handy as a teacher*; adapted from Gaspard et al., 2017), cost of the training ($\alpha = .901$, 3 items, sample item=*I have to give up a lot of things to do well in the courses of my training*; adapted from Gaspard et al., 2017), and college self-efficacy ($\alpha = .777$, 4 items, sample item=*Do well on your exams*; Solberg et al., 1993), all rated on a 6-point Likert-type scale of agreement. Covariates also included self-reported grade point average (GPA) (21-point scale).

2.3 Data analysis

2.3.1 Reliability

To evaluate the reliability of the LLM's encoding, one researcher (encoder 1) and one collaborator (encoder 2) independently encoded a subset of 277 randomly selected responses (comprising 129 on resources and 148 on demands) from a dataset of the previous academic year. This dataset was chosen because it contained responses to the same questions, aimed at measuring the same variables within the same population. This selection allowed us to test the reliability of the method with a similar sample size to the one used in this study while ensuring that the data used for the validity analyses presented in this paper were coded exclusively by the LLM. In this process, both the LLM and the human encoders independently categorized each response by counting the occurrences of each specific academic resource or demand mentioned. To quantify the reliability, we employed the Cohen's weighted kappa statistic, using JASP 0.18.1 (JASP Team, 2023) for the calculation. Cohen's kappa values ranging from 0.61 to 0.80 indicate substantial agreement, while values between 0.81 and 1 suggest almost perfect agreement (Landis & Koch, 1977).

2.3.2 Bivariate correlation and network analysis

We examined the validity of the automatic encoding procedure via LLM by assessing the similarity between LLM and scale measures of study demands and resources. For this, we used two indicators of validity: convergent and concurrent validity.

For convergent validity, we investigated the correlation between the two kinds of measures, while assuming that high correlations would indicate that the two methods enabled measuring the same latent constructs (Campbell & Fiske, 1959). However, prior studies have demonstrated that count data may not exhibit a strong correlation with scale data even when they are measuring the same constructs, due to the nature of the data types (Pluye et al., 2009).

To address this potential limitation, we supplemented our validation with a concurrent validity analysis. Specifically, we used network analysis as an alternative statistical test to examine whether LLM and scale measures are similarly related to other significant variables in our dataset. Network analysis allowed us to visualize and quantify the structural relationships and interconnections among various constructs relevant to teacher education (e.g., value and cost of training, peer support, self-efficacy, grade point average). This methodology enables the examination of the structural relation among different variables, facilitating the visualization and quantification of the connections' strength and patterns (Borsboom et al., 2021). A similarity in the structural configurations of both networks (i.e., with count vs. scale data) would suggest that the demands and resources variables, regardless of the measurement method, maintain consistent relations with other covariates. Such an observation would lend empirical support to the premise that these variables are effectively capturing the same processes.

2.3.3 Response surface analysis

Finally, to shed light on the information provided by LLM measures about the focal theory of study demands-resources fit (RQ2), we analyzed how the interplay between study demands and resources contributed to explain variations in student teachers' engagement. To this end, we used cubic response surface analysis (cubic RSA), based on a 3-step identification strategy (Núñez-Regueiro & Wang, 2025; Humberg et al., 2020). This technique allows predictor variables (e.g., $x = demands$, $y = resources$) to have additive, interactive, quadratic, cubic, and interactive-quadratic effects on the outcome variable (e.g., $z = student\ engagement$), by specifying a cubic polynomial model c ($z = b_1x + b_2y + b_3x^2 + b_4xy + b_5y^2 + b_6x^3 + b_7x^2y + b_8xy^2 + b_9y^3$). Therefore, RSA allows to capture and represent non-linear relationships in a three-dimensional space, including optimal fit margins (quadratic nonlinearities) with potential flow or relaxation effects (cubic nonlinearities; Fig. 1). By accounting for different types of non-linearities, cubic RSA provides a better understanding of how engagement varies across demands-resources (mis)fit conditions.

The first step allows us to determine whether the data are better explained by linear or quadratic surfaces, or by more complex curvatures involving cubic or interaction terms. This step involves selecting the best-fitting solution from among 37 families of polynomials, which are defined by applying specific zero or proportionality

constraints to first-order (b_1, b_2), second-order ($b_3,\ b_4, b_5$), or third-order polynomial parameters ($b_6,\ b_7, b_8, b_9$). Best-fitting solutions are selected based on tests of absolute fit against the saturated cubic model, on parsimony and explanatory power indices (i.e., AIC weights, adjusted R2), on the significance of the family-specific parameters, and on evidence of good fit to the data using the benchmark values of CFI≥0.95, TLI≥0.95, RMSEA≤0.06, SRMR≤0.08 (Hu & Bentler, 1999). Step 2 then identifies the best-fitting variant within the best-fitting family, by testing equality or zero constraints on lower-order polynomial parameters in that family using likelihood ratio tests between nested models. In other words, this step refines the selected family by comparing alternative parameter combinations, further improving model parsimony without compromising fit. In Step 3, the robustness of results from the model-selection processes conducted in Steps 1 and 2 is assessed. This is done by utilizing 1000 bootstrapped samples and ensuring that the obtained 95% CI of non-null parameters in the best-fitting variant does not contain the value zero. In the present study, the above 3-step identification strategy was conducted after standardizing variables to unit variance and zero means. In addition, a robustness check was conducted to verify the best-fitting solution replicated while using scale measures instead of LLM measures.

Key to the RSA is the ability to visualize these complex relations in a three-dimensional space, providing an understanding of how different combinations of predictors impact the outcome variable. In the graphical interpretation, we rely on the line of fit between demands and resources known as "line of congruence" (LOC, $demands = resources$) and the line of misfit or "incongruence" (LOIC, $demands = -resources$; Edwards, 1996). These lines are informative because they describe how the degree of fit or misfit relates to student teachers' engagement, such as when the levels of academic resources and demands are in harmony (e.g., high resources and high demands) or disharmony (e.g., low resources, high demands; see Fig. 1A). Using this approach, the best-fitting variant was interpreted by probing the curvatures of the response surface along the LOC and the LOIC, notably by considering reversal points "r" where the behavior changes valence („∂z-.-∂,x-. = 0.0; Núñez-Regueiro et al., 2025; Humberg et al., 2020).

2.3.4 Missing data treatment

Although missing data was low for scale data (<1% for all variables), it was more consequent for LLM data (39% for study demands, 56% for study resources). A comparison analysis showed that students whose count data was missing did not differ in terms of occurrences of demands or resources in count data, but they did report higher levels of demands and lower levels of resources in scale data (Supplemental Material A—SM-A). Moreover, this missing data was also related to social background characteristics (being older, reporting lower father and mother educational level, being in a couple) and to teacher education variables (reporting higher cost of training, lower value of training, lower self-efficacy; SM-A). Many other characteristics were not related to missing data (e.g., gender, father or mother SES, teaching level, perceived peer support, student teachers' engagement, grade point average),

suggesting the missingness mechanism was related to specific variables, an indication of missingness at random (MAR; Graham, 2012; Rubin, 1987).

An appropriate technique for handling missing data is using full information maximum likelihood (FIML), which takes into account all non-missing data to estimate model parameters (Graham, 2012). This strategy enabled accounting for 100% of the sample. Because missing count data was high, we also conducted a sensitivity analysis to see if the results changed in a nontrivial way after removing individuals with missing data (listwise deletion). As detailed in SM-B, this led to identical results as the FIML approach, indicating that the missing data treatment was not decisive in obtaining the final results.

3 Results

3.1 Reliability of automatic data encoding

Using Cohen's weighted kappa to measure the agreement between GPT and human encoders, as well as between human encoders themselves, the results (Table 1) demonstrate the reliability of the automatic encoding process. The findings revealed excellent average results for the classification of resources and substantial agreement for demands. Furthermore, it can be observed that LLM performance mirrored human performance: in the encoding of academic resources where humans excelled, the LLM also showed better performance; conversely, in the encoding of demands where humans had difficulties, the LLM exhibited weaker performance.

To further illustrate the results of Table 1, we provide a selection of examples of real responses encoded by the model, translated from French. These examples, chosen for being consistently classified by both human encoders, showcase the model's correct and incorrect categorization depending on whether they matched the human encodings. Examples of incorrect categorizations made by the model included overlooking or overcounting academic resources and demands in an answer. For study resources, a correctly classified answer was: "*The trainers encourage us*, *coach us*, *advise us and help us by constantly motivating us.*", which showcased the social support received from teacher educators, that is, a resource specific to teacher education. By contrast, an example of misclassified study resource answer was: "*My motiva-*

Table 1 Cohen's weighted kappa

Variable	Ratings	Weighted kappa
Resources	Average kappa across encoders and GPT	0.800
	Encoder 1 - GPT	0.806
	Encoder 1- Encoder 2	0.903
	GPT - Encoder 2	0.700
Demands	Average kappa across encoders and GPT	0.621
	Encoder 1 - GPT	0.662
	Encoder 1- Encoder 2	0.667
	GPT - Encoder 2	0.535

277 coded answers and 3 raters

tion comes from my work placement on Mondays. When I see the students, I tell myself that this is really what I want to do. The idea that I have to carry on so that I haven't done a 5-year degree for nothing. making your family proud. the support of your fellow students and your family, as well as the instructors at the [blinded] site, is very important.". This example was misclassified because the main motivator related to the teaching experiences provided by practical classes (work placement on Mondays), that is, a vocational resource, whereas the remaining were not clearly academic in nature.

Similarly, an example of correctly classified answers when assessing study demands was: "*Sometimes a lot of study work at the same time, which doesn't necessarily leave us the time to be 'all in' on each task.*", and an incorrect one was: "*There's a huge amount of personal work to be done, like reading a lot and doing summary sheets, making sheets on the lessons etc… It takes up a huge amount of time and gives you the impression that you'll never be able to finish anything. What's more, as I'm doing a training course for teaching physical education, I also have to take physical training into account to prepare for the recruitment contest, so it's difficult to manage both lessons and training, not to mention the daily tasks I have to do as well. All this work leads to overwork and a great deal of anxiety.*". Despite starting similarly to the first, the second response was much longer and contained a greater mixture of ideas, which likely contributed to its incorrect classification. These examples highlight that the misclassified responses were noticeably more intricate and lengthier compared to those correctly classified, which were simpler.

Following the reliability check of the data obtained through automatic encoding, we proceeded to encode the 524 answers to the open-ended questions used for data analysis in this study. The process took approximately 30 min. Table 2 showcases the results, detailing the frequency of academic resources and demands identified in the student teachers' responses.

Table 2 Frequency of academic resources and demands in student teachers' responses as identified by the automatic encoding

Variable	Number of different academic resources/demands present in the answer	Frequency	Percentage
Resources	0	138	62.44
	1	61	27.60
	2	15	6.79
	3	7	3.17
	Total	221	100
Demands	0	36	11.88
	1	139	45.87
	2	87	28.71
	3	26	8.58
	4	9	2.97
	5	6	1.98
	Total	303	100

3.2 Validity of automatic encoding

3.2.1 Correlation with scale measures (Convergent validity)

In assessing the convergent validity of our study, we investigated the correlations between data obtained from scales and the count data derived from the LLM's encoded answers to the open-ended questions. The correlation for measures of study demands was found to be $r_{study\ demands} = 0.114$, with a probability value of $p = .043$. On the other hand, the correlation for study resources yielded a stronger positive relationship, with $r_{study\ resources} = 0.205$ ($p < .001$). These significant correlations underscore a consistent, positive relation between the data obtained from scales and the LLM-encoded count data for both study demands and resources. Notwithstanding their significance, the correlations were small in terms of effect size (i.e., $0.10 \leq r \leq .30$; Cohen, 1988), pointing to low degrees of convergent validity.

3.2.2 Structural relations to other constructs (Concurrent validity)

The outcomes of the network analysis are depicted graphically below, facilitating a comparative examination of the network's structural configuration (see Fig. 3). This comparison highlights the differences and similarities in the more general correlational structure of teacher education variables, when employing demands and resources measures derived from scales versus LMM data. The examination of the structural configurations of the networks reveals consistency between LLM and scale measures, but with some variations. Both LLM and scale measures of study resources

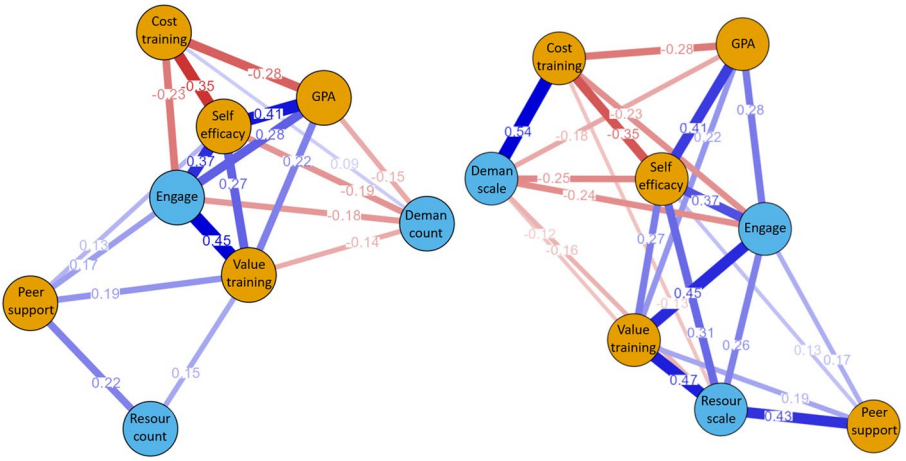

Fig. 3 Structural configuration of the network analysis. Note. The upper image depicts the structural configuration between the main variables (blue) and the covariates (orange), using LMM measures of demands and resources obtained through open-ended questions ("Deman count" and "Resour count"). Lower image illustrates a similar structural configuration but utilizes data sourced from scales ("Deman scale" and "Resour scale"). Only significant correlations are shown: blue edges for positive and red edges for negative correlations. Engage = Engagement; GPA = Grade point average; Resour = Resources; Deman = Demands

exhibited positive correlations with the perceived value of the training and the amount of peer support. Regarding study demands, LLM and scale data also yielded similar correlations, albeit with more subtle differences. For both types of data, demands were negatively related to self-efficacy and perceived value of training and positively related to the perceived costs of training (all relations significant at 5%). Nevertheless, the relation between demands and costs of training was much weaker with LLM data compared to scale data (r =.09 vs. r=.54, respectively). For all configurations, scale measures demonstrated stronger associations with adjacent variables, whereas count measures displayed weaker relations. Therefore, results of network analyses underscored the multidimensional concurrent validity of count and scale measures.

3.2.3 Response surface analyses of demands-resources fit and student teachers' engagement

Identification of best-fitting polynomial model After comparing the 37 polynomial families tested in the RSA, the best-fitting solution was the family of incongruence effect (FM8), characterized by zero constraints on third-order polynomials ($b_6 = \ldots = b_9 = 0$) and by equality constraints on second-order polynomials ($b_3 = \frac{b_4}{2} = b_5$). More specifically, this solution passed the test of absolute fit to the saturated cubic model ($\chi^2_6 = 2.7$, p =.849), ranked highest in model parsimony (wAIC = 0.212), and showed excellent fit to the data (e.g., CFI = 1.000, RMSEA < 0.001, SRMR = 0.006; Table 3). An inspection of parameters (Model 1, Table 4) suggested the main effects of demands ($b_1 = -0.128$) could be constrained to be 1.5 larger than the effect of resources ($b_2 = 0.085$; Model 1, Table 4). These constraints could be imposed without loss of generality ($\chi^2_1 < 0.001$, p =.984). This obtained a best-fitting variant for FM8 (Step 2; Model 2, Table 4) that explained 5% of the variance in student teachers' engagement. This represented a 92% increase in the explained variance compared to a model including only the additive effects of demands and resources (i.e., 2.6%; Table 3). Finally, bootstrapping procedures showed that the 95% CIs of parameters from this best-fitting variant did not contain zero values and therefore allowed for valid post-model-selection inferences (Model 3, Table 4).

Interpretation of response surface The response surface of the best-fitting variant aligned with the optimal margin of fit hypothesis (H2), characterized by a negative quadratic effect along the LOC (line of fit), and a negative trend effect along the LOIC (line of misfit; see Fig. 4). The optimal point was near average values at 0 (i.e., $demands = resources = -0.094$; Fig. 4B). Therefore, experiencing balanced levels of resources was conducive to high levels of engagement, but this adaptive effect waned out if the levels of demands and resources were below or above average. By contrast, experiencing more demands than resources was negatively and linearly related to engagement (i.e., $v_1 = b_1 - b_2 = -0.213$), indicating that engagement levels decreased when demands overwhelmed resources ($demands = -resources > 0$), and increased when resources outweighed demands ($demands = -resources < 0$; Fig. 4C).

Table 3 Fit indices of polynomial models of demands-resources processes of student teachers' engagement, ranked by parsimony

Model	χ^2	DF	*p*	wAIC	R2	CFI	TLI	RMSEA	SRMR
Cubic model (saturated model)	0.0	0	—	0.002	0.047	1.000	1.000	0.000	0.000
8. Incongruence effect	**2.7**	**6**	**0.849**	**0.212**	**0.048**	**1.000**	**1.366**	**0.000**	**0.006**
Best-fitting variant of 8	***2.7***	***7***	***0.914***	**—**	***0.050***	***1.000***	***1.408***	***0.000***	***0.006***
11. Incongruence effect curved by X	3.7	6	0.717	0.127	0.043	1.000	1.252	0.000	0.010
14. Rotated congruence effect	2.7	5	0.751	0.078	0.046	1.000	1.307	0.000	0.006
12. Incongruence effect curved by Y	4.9	6	0.554	0.069	0.042	1.000	1.119	0.000	0.013
1. Main effect of X	9.4	8	0.307	0.053	0.028	0.895	0.882	0.019	0.021
4. Interaction effect between X and Y	5.8	6	0.440	0.043	0.039	1.000	1.016	0.000	0.015
25. Non-parallel, asymmetric congruence and incongruence effects	0.4	3	0.935	0.032	0.051	1.000	1.566	0.000	0.002
5. Quadratic effect of X	6.6	6	0.363	0.030	0.032	0.958	0.937	0.014	0.016
28. Non-parallel, asymmetric weak congruence and strong incongruence effects	0.8	3	0.846	0.027	0.050	1.000	1.480	0.000	0.002
18. Level-dependent quadratic effect of X	1.1	3	0.775	0.023	0.048	1.000	1.415	0.000	0.002
3. Additive effects of X and Y	9.4	7	0.223	0.020	0.026	0.822	0.771	0.026	0.021
13. Quadratic effects of X and Y	5.4	5	0.364	0.019	0.035	0.967	0.941	0.013	0.014
20. Asymmetric congruence effect	1.5	3	0.685	0.019	0.047	1.000	1.332	0.000	0.005
21. Asymmetric incongruence effect	1.6	3	0.652	0.018	0.046	1.000	1.301	0.000	0.004
23. Level-dependent incongruence effect	1.8	3	0.613	0.016	0.045	1.000	1.261	0.000	0.007

X = demands; Y = resources. In plain characters, numbered models correspond to the 15 best-fitting families of fundamental models of congruence, which are compared in Step 1 of the identification strategy (for a list of all 37 families, see SM-B, Table S2). Polynomial models in bold and italic characters correspond to the best-fitting family (Step 1) and the best-fitting variant within the family (Step 2), respectively

Table 4 Parameters of best-fitting model of student teachers' engagement (3-Step Identification)

Model	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9
Model 1: Best-fitting polynomial family FM8 (Step 1)									
Estimate	−0.129	0.084	−0.056	−0.112	−0.056	0	0	0	0
p-value	0.061	0.276	0.015	0.015	0.015	(fixed)	(fixed)	(fixed)	(fixed)
Model 2: Best-fitting variant of FM8 (Step 2)									
Estimate	−0.128	0.085	−0.056	−0.113	−0.056	0	0	0	0
p-value	0.032	0.032	0.010	0.010	0.010	(fixed)	(fixed)	(fixed)	(fixed)
Model 3: Best-fitting variant of FM8, bootstrapped (Step 3)									
Estimate	−0.169	0.113	−0.078	−0.156	−0.078	0	0	0	0
CI 95%	(−0.300, −0.054)	(0.036, 0.200)	(−0.138, −0.016)	(−0.277, −0.032)	(−0.138, −0.016)	(fixed)	(fixed)	(fixed)	(fixed)

N=499 student teachers. Parameters b_1 to b_9 describe the cubic polynomial effects of study demands (x) and resources (y) on student teachers' engagement in teacher education (z; see Fig. 1B). Zero values are fixed by construction, as part of the identification strategy of the best-fitting polynomial. Bootstrapped values correspond to means and 95% confidence intervals (1000 samples). The response surface of the best-fitting variant is reported in Fig. 4

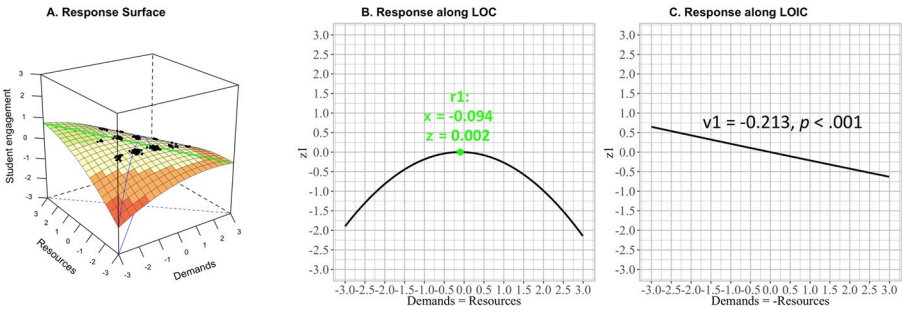

Fig. 4 Response surface of demands-resources fit. Note. N=499 students in teacher education. LLM=large language models; LOC=line of congruence ($demands = resources$); LOIC=line of incongruence ($demands = -resources$). Lines in green intersect reversal points on the LOC, indicating an optimal margin of demands-resources fit. Model parameters are reported in Table 3 (Model 2)

Robustness check with scale measures As reported in the Supplemental Material, the identified response surface replicated after substituting LLM measures with scale measures (see Table S3 and Figure S2, SM-B). The replication concerned both the identification of the best-fitting family, the pattern of optimal margin in demands-resources fit (H2), and the location of the reversal point along the LOC (i.e., $demands = resources = 0.000$; Figure S2). The only major difference concerned the size of this pattern, which showed slightly weaker curvature along the LOC and slightly stronger negative trend along the LOIC. However, this difference in size did not modify the substantive interpretation of the surface. In summary, the solution identified through LLM measures appeared to be robust to more traditional measurement approaches (i.e., scale measures).

4 Discussion

In this study, we aimed to evaluate the reliability and validity of LLM-derived data for statistical analysis (methodological focus, RQ1) and their heuristic value in revealing how patterns of demands-resources fit relate to student teachers' engagement (substantive focus, RQ2). Overall, our findings highlight the added value of a methodological-substantive synergy (Marsh, 2007; Marsh & Hau, 2007), demonstrating that LLM-encoded open-ended data can serve as a reliable and valid alternative to traditional structured measures, while also enriching our understanding of how demands and resources shape engagement in teacher education. These results have important implications for research practices and theory on D-R fit processes in teacher education and are discussed in detail in the sections that follow.

4.1 Large language models can produce valid measures from open-ended data

The first contribution of this study was to assess the reliability and validity of LLM in encoding open-ended data on study demands and resources, as reported in writing by students in teacher education (RQ1). This was evidenced in two major ways.

First, LLM was shown to be as reliable as human encoders, presenting similar difficulties in encoding participant answers to open-ended questions. Two human encoders and LLMs independently evaluated 277 randomly selected texts, finding good and excellent agreement, respectively, for encoding study demands and resources (Cohen's weighted kappa). These findings highlight the reliability of LLMs in retrieving information from open-ended question responses. Moreover, misclassified responses by the LLM were significantly more complex than correctly classified ones, causing the LMM to overlook or overcount demands or resources compared to humans. However, this misclassification pattern was also observed among human encoders, indicating that text complexity challenges both human and machine encoders (Krippendorff, 2004; Strijbos et al., 2006). Additionally, responses about study demands were generally longer and more disorganized than those about study resources, showing a tendency among student teachers to describe more in-depth negative experiences more than positive ones (negativity bias phenomenon; Amabile & Glazebrook, 1982; Kim et al., 2022; Poncheri et al., 2008). This contributed to more discrepancies among both human and LLM encoders in determining the number of demands in lengthy responses, leading to lower kappa values. These insights into the obstacles created by overly complex or lengthy participant writings have implications for designing open-ended questions in future studies (see Implications).

Second, LLM measures provided similar information about individual differences in study demands and resources as traditional Likert-type scales. Both measures correlated systematically with each other and showed comparable relations to other variables (e.g., value and cost of teacher education, college self-efficacy, perceived peer support, GPA). Notwithstanding their similar structural relations to other variables, the correlation between LLM and scale measures was weak. This may be due to the specificity of the scales, which targeted particular sources of study demands (i.e., study workload) and resources (i.e., teacher need-supportive behavior). In contrast, responses to open-ended questions were more varied, including other sources of

stress or motivation not covered by scales, such as the perceived relevance of teacher education courses (Núñez-Regueiro et al., 2024; Deasy et al., 2016; Flores & Niklasson, 2014), or intrinsic motivation for teaching (Núñez-Regueiro et al., 2024; E. Kim & Corcoran, 2018; Watt & Richardson, 2007).

In addition to this, there are other possible explanations for the observed correlation values. One of them could be the nature of the data. LLM-derived measures are based on counts of spontaneously mentioned content, whereas scale measures are composed of structured, pre-defined items with interval response formats. These methodological differences may affect the precision and comparability of responses across individuals (Pluye et al., 2009). Finally, another factor that may have affected the correlation values was the different measurement errors. Both the encoding process and the scales had less than perfect reliabilities, contributing to some loss of information and affecting their correlation (Bland & Altman, 1996). Taken together, the observed correlations are conceptually expected and do not undermine the methodological value of LLM-derived measures because, despite differences in content, data type, and measurement error, LLM and scale measures exhibited similar relations to other covariates, indicating strong concurrent validity. They also revealed the same process of demands-resources fit in relation to student teachers' engagement in teacher education. Consequently, it might be concluded that, although based on spontaneous responses to open-ended questions, LLM measures enabled obtaining statistical results coherent with those obtained from scale measures, indicating structural validity. This underscores the relevance of future studies using LLM measures for applied statistics with qualitative data.

Considering these findings collectively, LLM-based analyses could help bridge gaps in traditional assessment methods. For example, integrating LLM-driven text analysis into digital learning platforms could enable automatic categorization of student responses, improving adaptive learning models (Falcon & Leon, 2024; Demszky et al., 2025; Demszky et al., 2023a, b). This could allow educational systems to identify students' specific academic challenges and tailor interventions accordingly, supporting both personalized learning experiences and large-scale educational assessment frameworks.

4.2 Student engagement in teacher education responds to an "optimal margin" of demands-resources fit

The second contribution of this study was shedding light on student teachers' engagement processes in teacher education, based on LLM measures of demands-resources fit processes (RQ2). Aligning with the optimal margin of D-R fit (Fig. 1D), results showed that study resources can mitigate the impact of study demands, but only to a certain limit (Núñez-Regueiro, 2017; Núñez-Regueiro et al., 2022; Dupéré et al., 2015). As increasingly high demands were met with increasingly high resources, student teachers' engagement appeared to decrease. Conversely, too few demands met by too few resources were also counterproductive. According to the D-R fit circumplex (Fig. 1A), this pattern of optimal D-R fit indicates that when demands activate a commensurate amount of resources, too little activation may lead to boredom (negative fit), whereas too much activation may lead to burnout (positive fit; Núñez-Regueiro,

2017; Núñez-Regueiro et al., 2025). Only intermediate levels (in this study, average levels) of commensurate demands and resources may be optimal. This quadratic non-linearity is similar to quadratic relations of stimuli to learning performance or arousal response (Yerkes & Dodson, 1908), but in D-R fit theory, it relates to commensurate amounts of demands and resources. By contrast, when amounts are not commensurate (D-R misfit), a relative excess in resources (high resources before low demands) was systematically and linearly related to high student teachers' engagement levels.

Taken together, these findings contribute to a more nuanced understanding of the interplay between student demands and resources in educational settings. First, accounting for the interplay between demands and resources (i.e., optimal margin of D-R fit) nearly doubled the predictive power of the model (+92%), suggesting that D-R fit processes are highly complex. This finding aligns with individual stress theories (Núñez-Regueiro, 2017; Núñez-Regueiro et al., 2022; Núñez-Regueiro & Leroy, 2024; Edwards, 1996; Lazarus, 1999; Pearlin, 1999), and confirms the need to move the D-R fit theory from an additive framework (Demerouti et al., 2001; Salmela-Aro & Upadyaya, 2014) to an interactive framework (Bakker & Demerouti, 2017; Salmela-Aro et al., 2022). Furthermore, it qualifies previous findings concluding to the buffering effect of resources against study demands (Núñez-Regueiro, 2018; Núñez-Regueiro et al., 2022; Núñez-Regueiro & Leroy 2024; Yang et al., 2018). The present study as well as preliminary findings (Núñez-Regueiro et al., 2025) now converge to say that resources can mitigate the negative impact of study demands only to a limit (optimal margin), or only when they actually exceed the level of demands (positive misfit).

4.3 Limitations and research perspectives

This study utilized a commercial LLM due to the convenience offered by the company in analyzing extensive datasets. However, this approach implied financial costs (albeit rather low). As the field of natural language processing rapidly evolves, new open-source LLMs are emerging, such as Llama, Qwen or Mistral. These models present interesting perspectives for enhancing the automatic encoding of open-ended responses at no economic cost, making them more accessible to the scientific community. Future studies should examine the performance of these models in comparison to the commercial approach used here, with the goal of identifying the most effective and efficient tools. In addition, further research should also investigate how these models operate across diverse educational contexts, evaluate their cross-linguistic applicability, and refine prompt design to improve output consistency and interpretability. Together, these efforts will contribute to the responsible and equitable integration of LLMs in educational research.

Another limitation of our study was the demonstration of predictive validity using only a single sample and for a specific process. To validate the technique more robustly, future research should employ more diversified research designs and questions. Such expanded exploration is crucial for establishing the generalizability and applicability of these methods across different contexts and research scenarios (Scandura & Williams, 2000). In addition, future studies should also examine whether

LLM-based measures perform similarly across different theoretical frameworks, beyond the D-R fit model used here.

An additional consideration is the role of content analysis and interpretive analysis. Often justified due to the challenges of integrating large samples into qualitative research, these methodologies remain essential. The advent of tools like LLMs allows us to overcome some of these challenges, particularly in content quantification. However, it is important to acknowledge that preliminary research is necessary to provide guidelines for generating instructions for these tools. Therefore, while LLMs and similar technologies are not replacements for qualitative data analysis, they serve as valuable extensions, enabling us to derive more generalizable findings from large datasets (Demszky et al., 2023).

4.4 Implications for research and theory

The findings of this study have significant implications for both methodology and theory. Methodologically, our work demonstrates the usefulness of LLMs for automatically encoding responses to open-ended questions. This advancement enables the reliable collection and analysis of large datasets quickly, facilitating the use of open-ended questions to assess various variables in different contexts. For example, Núñez-Regueiro et al. (2024) manually encoded 160 texts on resources and demands, taking 3 to 4 full weeks per encoder, with three encoders involved (a total of 3 months of combined effort). In contrast, our study used an LLM to process data from hundreds of individuals in less than an hour. Although exact time savings are hard to quantify due to variables like server status and encoder availability, the improvement is substantial. This significant reduction in time and effort highlights the benefits of automatic encoding (Kaufmann et al., 2020; Leeson et al., 2019; Rietz & Maedche, 2021). Furthermore, our study showed better reliability indices compared to the manual approach in Núñez-Regueiro et al. (2024), where reliability scores for the two dimensions were lower. This could be due to decision fatigue experienced by human encoders working with large volumes of text over extended periods, affecting their consistency and accuracy (Vohs et al., 2005). These results support the utility of automatic coding with LLMs, benefiting researchers seeking insights on specific variables for diverse analyses using open-ended questions, while also enabling advanced statistical analysis with large samples.

Another methodological contribution of this study is the approach to formulating open-ended questions. As observed, some participants tended to write excessively or stray off-topic, which resulted in misclassified encodings (for both humans and the LLM). To counter this, we suggest that such questions, intended for either human or LLM analysis, should follow guidelines promoting concise, structured, and clear responses to ensure more precise encoding. For example, requesting respondents to enumerate ideas rather than offering free-form text may limit the response spontaneity, but yield data more amenable to encoding. The challenge lies in striking an optimal balance between the spontaneity of responses and their semi-structured presentation.

These findings underscore the transformative potential of LLMs in educational assessment, particularly in enhancing digital assessment tools and large-scale ana-

lytics. By demonstrating that LLM-derived measures are both reliable and valid, this study highlights their potential to complement traditional structured data collection methods, offering a scalable approach to analyzing open-ended responses. This methodological advancement can support LLM-powered educational platforms by enabling automated and personalized feedback, identifying learning patterns, and optimizing student support systems (Falcon & Leon, 2024; Demszky et al., 2025; Demszky et al., 2023a).

Nonetheless, the adoption of LLM-based approaches also entails practical challenges for educational researchers. These include the need for technical expertise to properly formulate prompts and interpret model outputs, ensuring transparency in the classification process, and addressing potential biases embedded in the language models (Bai et al., 2025; Lee et al., 2024). Overcoming these barriers will require the development of clearer methodological guidelines, accessible tools, and interdisciplinary collaboration between education researchers and AI developers. In addition, ethical and legal concerns, particularly regarding data privacy and confidentiality, are central when using commercial LLMs that process data through external servers. To mitigate these risks, we highlight the potential of open-source LLMs (e.g., Llama, Qwen, Mistral, etc.) that can be deployed on local infrastructures, allowing for on-premise processing of sensitive data.

On the other side, the results from the automatic encoding have also contributed from a substantive point of view. The encoded data revealed a higher prevalence of academic demands than resources among student teachers: out of 303 responses, 267 indicated at least one demand, compared to 83 out of 276 responses indicating at least one academic resource. This finding aligns with prior research (Núñez-Regueiro et al., 2024), which observed that student teachers are generally experiencing greater stress due to academic factors. The greater stress due to academic factors arises from challenges in the teacher education program, such as irrelevant coursework, heavy workload, and a gap between educational content and practical teaching skills. Furthermore, building on the present study, we can assert that resources can only mitigate the negative impact of study demands up to a certain threshold (optimal margin) or when they surpass the level of demands (positive misfit). This insight carries practical implications, suggesting that educational policies and curriculum designs that alleviate demands, or enhance resources to effectively balance or surpass academic demands, could significantly alleviate student teachers' stress and improve educational outcomes. For example, study demands could be reduced by reorganizing the curriculum so as to space out assignments and study workload (Núñez-Regueiro et al., 2024), but also by increasing the perceived relevance of courses to teaching (Clarke et al., 2012; Deasy et al., 2016) and fostering hands-on experiencing with teaching by alternating school- and college-based training (Hennissen et al., 2017). Complementarily, study resources could be enhanced by fostering positive relations with teacher educators and college peers to provide emotional support (Kim & Corcoran, 2018) and developing problem-solving coping strategies among student teachers (Gustems-Carnicer & Calderón, 2013).

5 Conclusions

This study marks a significant methodological and theoretical advancement in using LLMs for the automatic encoding of open-ended responses, with a particular focus on student teachers' academic resources and demands. Methodologically, one of the standout achievements of this study is the demonstration of LLMs' efficiency in data encoding. The ability to encode a large amount of open-ended data not only saves time but also opens new possibilities for conducting research that was previously impractical. This efficiency does not come at the cost of depth; rather, it enables a different type of analysis of large datasets, making it a valuable tool for researchers across various disciplines.

From a theoretical perspective, our findings offer insights into the dynamics of student teachers' academic resources and demands, and engagement. The data indicates that resources can effectively mitigate the negative impact of study demands but only to a certain extent, or when they surpass the level of demands. This insight is crucial for understanding how the balance between resources and demands shapes student teachers' stress levels and engagement. Practically, these findings have significant implications for educational policy and curriculum design. They suggest that strategically reducing study demands or enhancing student resources to either match or exceed demands could considerably reduce student teachers' stress and enhance educational outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10639-025-13764-6.

Authors contribution Fernando Núñez-Regueiro: Conceptualization, Investigation, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing. Samuel Falcon: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing. Pascal Bressoux: Conceptualization, Writing - Review & Editing, Supervision.

Funding This work was supported by University of Las Palmas de Gran Canaria, Cabildo de Gran Canaria, and Banco Santander through the pre-doctoral training programme for research personnel.

Data availability Data is available under request.

Declarations

Ethical approval and informed consent statements.

Conflict of interest The authors report there are no competing interests to declare.

References

Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., & Talpur, N. (2023). Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges. *IEEE Access : Practical Innovations, Open Solutions, 11*, 10535–10562. https://doi.org/10.1109/ACCESS.2023.3240898

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

Alqassab, M., & Leon, J. (2024). Motivational messages from teachers before exams: Links to intrinsic motivation, engagement, and academic performance. *Teaching and Teacher Education*. https://doi.org/10.1016/j.tate.2024.104750

Amabile, T. M., & Glazebrook, A. H. (1982). A negativity bias in interpersonal evaluation. *Journal of Experimental Social Psychology*, *18*(1), 1–22. https://doi.org/10.1016/0022-1031(82)90078-6

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*. https://doi.org/10.1016/j.econedurev.2019.101919

Bailey, J. (2008). First steps in qualitative data analysis: Transcribing. *Family Practice*, *25*(2), 127–131. https://doi.org/10.1093/fampra/cmn003

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.2416228122

Bakker, A. B., & Demerouti, E. (2017). Job demands-resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, *22*(3), 273–285. https://doi.org/10.1037/ocp0000056

Bland, J. M., & Altman, D. G. (1996). Measurement error and correlation coefficients. *BMJ (Clinical Research Ed.), 313*(7048), 41–42. https://doi.org/10.1136/bmj.313.7048.41

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A. M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*. https://doi.org/10.1038/s43586-021-00055-w

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*. https://doi.org/10.48550/arXiv.2005.14165

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, R., Goodman-Williams, R., Feeney, H., & Fehler-Cabral, G. (2020). Assessing triangulation across methodologies, methods, and stakeholder groups: The joys, woes, and politics of interpreting convergent and divergent data. *American Journal of Evaluation*, *41*(1), 125–144. https://doi.org/10.1177/1098214018804195

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating mixed methods research with natural Language processing of big text data. *Journal of Mixed Methods Research*, *15*(3), 398–412. https://doi.org/10.1177/15586898211021196

Chiang, C. H., & Lee, H. Y. (2023). Can large language models be an alternative to human evaluation? *Annual Meeting of the Association for Computational Linguistics, 1*, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

Choi, J. N. (2004). Person-environment fit and creative behavior: Differential impacts of supplies-values and demands-abilities versions of fit. *Human Relations, 57*(5), 531–552. https://doi.org/10.1177/0018726704044308

Clarke, M., Lodge, A., & Shevlin, M. (2012). Evaluating initial teacher education programmes: Perspectives from the Republic of Ireland. *Teaching and Teacher Education*, *28*(2), 141–153. https://doi.org/10.1016/j.tate.2011.08.004

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203771587

Creswell, J. W., & Creswell, J. D. (2017). Qualitative procedures. In *Research design: Qualitative, quantitative, and mixed methods approaches* (pp. 162–187).

Deasy, C., Coughlan, B., Pironom, J., Jourdan, D., & Mannix-McNamara, P. (2016). Psychological distress and help seeking amongst higher education students: Findings from a mixed method study of undergraduate nursing/midwifery and teacher education students in Ireland. *Irish Educational Studies*, *35*(2), 175–194. https://doi.org/10.1080/03323315.2016.1146157

Demerouti, E., Nachreiner, F., Bakker, A. B., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology*, *86*(3), 499–512. https://doi.org/10.1037/0021-9010.86.3.499

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023a). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis,* , Article 016237372311692. https://doi.org/10.3102/01623737231169270

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C. A., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023b). Using large Language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2025). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers and Education*. https://doi.org/10.1016/j.compedu.2024.105183

DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). Sage.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American chapter of the association for computational linguistics*. https://api.semanticscholar.org/CorpusID:52967399

Dupéré, V., Leventhal, T., Dion, E., Crosnoe, R., Archambault, I., & Janosz, M. (2015). Stressors and turning points in high school and dropout: A stress process, life course framework. *Review Of Educational Research, 85*(4), 591–629. https://doi.org/10.3102/0034654314559845

Edwards, J. R. (1996). An examination of competing versions of the person-environment fit approach to stress. *Academy of Management Journal, 39*(2), 292–339.

Falcon, S., & Leon, J. (2023). How do teachers engaging messages affect students? A sentiment analysis. *Educational Technology Research and Development*, *71*, 1503–1523. https://doi.org/10.1007/s11423-023-10230-3

Falcon, S., & Leon, J. (2024). Towards an optimised evaluation of teachers' discourse: The case of engaging messages. *arXiv*. https://doi.org/10.48550/arXiv.2412.14011

Flores, M. A., & Niklasson, L. (2014). Why do student teachers enrol for a teaching degree? A study of teacher recruitment in Portugal and Sweden. *Journal of Education for Teaching*, *40*(4), 328–343. https://doi.org/10.1080/02607476.2014.929883

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*(1), 148–162. https://doi.org/10.1037/0022-0663.95.1.148

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, *48*, 67–84. https://doi.org/10.1016/j.cedpsych.2016.09.003

Glazier, R. A., Boydstun, A. E., & Feezell, J. T. (2021). Self-coding: A method to assess semantic validity and bias when coding open-ended responses. *Research and Politics*. https://doi.org/10.1177/20531680211031752

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-4018-5

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S.Linden, S., & Van Der (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods, 29*(3), 494–51 https://doi.org/10.1037/met0000540

Gustems-Carnicer, J., & Calderón, C. (2013). Coping strategies and psychological well-being among teacher education students: Coping and well-being in students. *European Journal of Psychology of Education, 28*(4), 1127–1140. https://doi.org/10.1007/s10212-012-0158-x

Hennissen, P., Beckers, H., & Moerkerke, G. (2017). Linking practice to theory in teacher education: A growth in cognitive structures. *Teaching and Teacher Education, 63*, 314–325. https://doi.org/10.1016/j.tate.2017.01.008

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Humberg, S., Schönbrodt, F. D., Back, M. D., & Nestler, S. (2020). Cubic response surface analysis: Investigating asymmetric and level-dependent congruence effects with third-order polynomial models. *Psychological Methods*. https://doi.org/10.1037/met0000352

Hynninen, T., Knutas, A., Hujala, M., & Arminen, H. (2019). Distinguishing the themes emerging from masses of open student feedback. *2019 42nd International Convention on Information and Communication Technology Electronics and Microelectronics MIPRO 2019 - Proceedings*, *557–561*. https://doi.org/10.23919/MIPRO.2019.8756781

Jackson, J. C., Watts, J., List, J. M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science, 17*(3), 805–826. https://doi.org/10.1177/17456916211004899

JASP Team. (2023). *JASP (Version 0.18.1)*. https://jasp-stats.org/

Johnson, D. W., & Norem-Hebeisen, A. A. (1979). A measure of cooperative, competitive, and individualistic attitudes. *Journal of Social Psychology*, *109*(2), 253–261. https://doi.org/10.1080/00224545.1979.9924201

Johnson, R. B., & Onwuegbuzie, A. J. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*(5), 758–773. https://doi.org/10.1080/03075079.2011.598505

Kandel, B. (2020). Qualitative versus quantitative research. *Journal of Product Innovation Management*, *32*(5), 658.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kaufmann, A., Barcomb, A., & Riehle, D. (2020). Supporting interview analysis with autocoding. *Hawaii International Conference on System Sciences*, 752–761. https://doi.org/10.24251/HICSS.2020.094

Kim, D., Jung, W., Nam, S., Jeon, H., Baek, J., & Zhu, Y. (2022). Understanding information behavior of South Korean Twitter users who express suicidality on Twitter. *Digital Health*. https://doi.org/10.1177/20552076221086339

Kim, E., & Corcoran, R. P. (2018). Factors that influence pre-service teachers' persistence. *Teaching and Teacher Education, 70*, 204–214. https://doi.org/10.1016/j.tate.2017.11.015

Kim, H., Sefcik, J. S., & Bradway, C. (2017). Characteristics of qualitative descriptive studies: A systematic review. *Research in Nursing & Health, 40*(1), 23–42. https://doi.org/10.1002/nur.21768

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. https://doi.org/10.4135/9781071878781

Kuckartz, U. (2014). Qualitative text analysis using computer assistance. In *Qualitative text analysis: A guide to methods, practice & using software* (pp. 121–150). https://doi.org/10.4135/9781446288719

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. https://doi.org/10.2307/2529310

Lazarus, R. S. (1999). *Stress and emotion: A new synthesis*. Springer Publishing Co.

LeCompte, M. D., & Dworkin, A. G. (1991). *Giving up on school: Student dropouts and teacher burnouts*. Corwin Press, Inc.

Leech, N. L., & Onwuegbuzie, A. J. (2008). Qualitative data analysis: A compendium of techniques and a framework for selection for school psychology research and beyond. *School Psychology Quarterly*, *23*(4), 587–604. https://doi.org/10.1037/1045-3830.23.4.587

Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. In *British Journal of Educational Technology* (Vol. 55, Issue 5, pp. 1982–2002). John Wiley and Sons Inc. https://doi.org/10.1111/bjet.13505

Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural language processing (NLP) in qualitative public health research: A proof of concept sudy. *International Journal of Qualitative Methods, 18*, 1–9. https://doi.org/10.1177/1609406919887021

Le Mens, G., Kovács, B. I., Hannan, I. D., M. T., & Pros, G. I. (2023). *Uncovering the semantics of concepts using GPT-4*. https://doi.org/10.1073/pnas

Lennon, R. P., Fraleigh, R., van Scoy, L. J., Keshaviah, A., Hu, X. C., Snyder, B. L., Miller, E. L., Calo, W. A., Zgierska, A. E., & Griffin, C. (2021). Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family Medicine and Community Health*. https://doi.org/10.1136/fmch-2021-001287

Löfström, E., Poom-Valickis, K., Hannula, M. S., & Mathews, S. R. (2010). Supporting emerging teacher identities: Can we identify teacher potential among students? *European Journal of Teacher Education, 33*(2), 167–184. https://doi.org/10.1080/02619761003631831

Álvarez-Álvarez, C., & Falcon, S. (2023). Students' preferences with university teaching practices: Analysis of testimonials with artificial intelligence. *Educational Technology Research and Development*, *71*, 1709–1724. https://doi.org/10.1007/s11423-023-10239-8

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sample size matter in qualitative research? A review of qualitative interviews in is research. *Journal of Computer Information Systems*, *54*(1), 11–22. https://doi.org/10.1080/08874417.2013.11645667

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775–790. https://doi.org/10.1037/0022-0663.99.4.775

Marsh, H. W., & Hau, K. T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology*, *32*(1), 151–170. https://doi.org/10.1016/j.cedpsych.2006.10.008

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. https://doi.org/10.1080/00273170903333665

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2023). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics, 11*(1). https://doi.org/10.1177/20531680241231468

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification. *ACM Computing Surveys*. https://doi.org/10.1145/3439726

Mohajan, H. K. (2018). Qualitative research methodology in social sciences and related subjects. *Journal of Economic Development, Environment and People, 7*(1), 23. https://doi.org/10.26458/jedep.v7i1.571

Mu, J., Stegmann, K., Mayfield, E., Rosé, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*, *7*(2), 285–305. https://doi.org/10.1007/s11412-012-9147-y

Núñez-Regueiro, F. (2017). Dropping out of school as a stress process: Heterogeneous profiles in the form of boredom and burnout. *Orientation Scolaire Et Professionnelle*, *46*(1), 121–148. https://doi.org/10.4000/osp.5353

Núñez-Regueiro, F. (2018). *High school dropout: Analyzing the effects of stress processes and tracking, and dropout profile*. [Doctoral dissertation]. Université Grenoble Alpes.

Núñez-Regueiro, F., & Núñez-Regueiro, S. (2021). Identifying salient stressors of adolescence: A systematic review and content analysis. *Journal of Youth and Adolescence*, *50*(12), 2533–2556. https://doi.org/10.1007/s10964-021-01492-2

Núñez-Regueiro, F., & Juhel, J. (2022). Model-building strategies in response surface analysis. https://doi.org/10.17605/OSF.IO/YM32C

Núñez-Regueiro, F., & Juhel, J. (2024). *RSAtools: Advanced response surface analysis* (Version 0.1.1) [R CRAN].

Núñez-Regueiro, F., & Leroy, N. (2024). Exploring the joint effects of stressors and resources on student engagement in teacher education: A French study. *Teaching Education*. https://doi.org/10.1080/10476210.2023.2286356

Núñez-Regueiro, F., & Wang, M. T. (2025). Adolescent well-being and school engagement as a function of teacher and peer relatedness: The more (relatedness) is not always the merrier. *Journal of Educational Psychology, 117*(3), 466–484. https://doi.org/10.1037/edu0000910

Núñez-Regueiro, F., Archambault, I., Bressoux, P., & Nurra, C. (2021). Measuring stressors among adolescents: Validation of the positive and negative adolescent life experiences scale. *Journal of Psychoeducational Assessment, 39*(8), 969–982. https://psycnet.apa.org/doi/10.1177/07342829211027751

Núñez-Regueiro, F., Jamain, L., Laurent-Chevalier, M., & Nakhili, N. (2022). School engagement in times of confinement: A stress process approach. *Journal of Youth and Adolescence*, *51*(7), 1257–1272. https://doi.org/10.1007/s10964-022-01621-5

Núñez-Regueiro, F., Escriva-Boulley, G., Azouaghe, S., Leroy, N., & Núñez-Regueiro, S. (2024). Motivated to teach, but stressed out by teacher education: A content analysis of self-reported sources of stress and motivation among preservice teachers. *Journal of Teacher Education*. https://doi.org/10.1177/00224871231181374

Núñez-Regueiro, F., Vansoeterstede, A., Cappe, É., Boujut, E., Juhel, J., & Tang, X. (2025). Optimal margins in demands–resources fit and student engagement. https://hal.science/hal-05252118

OECD. (2019). *TALIS 2018 results (Volume I): Teachers and school leaders as lifelong learners*. TALIS, OECD Publishing.

Onwuegbuzie, A. J., & Daniel, L. G. (2003). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education*, *6*(2). https://cie.asu.edu/ojs/index.php/cieatasu/article/view/1609

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. T. (2012). Qualitative analysis techniques for the review of the literature. *The Qualitative Report, 17*, 1–28. https://doi.org/10.46743/2160-3715/2012.1754

OpenAI (2024). *FAQ - OpenAI Documentation*. Https://Platform.Openai.Com/Docs/Guides/Text-Generation/Faq. https://platform.openai.com/docs/guides/text-generation/faq

Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual Facebook posts. *Scientific Reports*. https://doi.org/10.1038/s41598-020-73917-0

O'Reilly, C. A., Chatman, J., & Caldwell, D. F. (1991). People and organizational culture: A profile comparison approach to assessing person-organization fit. *Management Journal*, *34*(3), 487–516. https://doi.org/10.2307/256404

Östlund, U., Kidd, L., Wengström, Y., & Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: A methodological review. *International Journal of Nursing Studies*, *48*(3), 369–383. https://doi.org/10.1016/j.ijnurstu.2010.10.005

Pearlin, L. I. (1999). The stress process revisited: Reflections on concepts and their interrelationships. In C. S. Aneshensel, & J. C. Phelan (Eds.), *Handbook of the sociology of mental health* (pp. 395–415). Kluwer/Plenum.

Pluye, P., Grad, R. M., Levine, A., & Nicolau, B. (2009). Understanding divergence of quantitative and qualitative data (or results) in mixed methods studies. *International Journal of Multiple Research Approaches*, *3*(1), 58–72. https://doi.org/10.5172/mra.455.3.1.58

Poncheri, R. M., Lindberg, J. T., Thompson, L. F., & Surface, E. A. (2008). A comment on employee surveys: Negativity bias in open-ended responses. *Organizational Research Methods*, *11*(3), 614–630. https://doi.org/10.1177/1094428106295504

Rahman, M. S. (2016). The advantages and disadvantages of using qualitative and quantitative approaches and methods in language testing and assessment research: A literature review. *Journal of Education and Learning, 6*(1), Article 102. https://doi.org/10.5539/jel.v6n1p102

Raj, H., Gupta, V., Rosati, D., & Majumdar, S. (2023). Semantic consistency for assuring reliability of large language models. *arXiv*. http://arxiv.org/abs/2308.09138

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C., Van Bavel, J. J., & Wood, J. (2023). GPT is an effective tool for multilingual psychological text analysis. *OSF*. https://doi.org/10.31234/osf.io/sekf5

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Rietz, T., & Maedche, A. (2021, May 6). Cody: An ai-based system to semi-automate coding for qualitative research. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445591

Roberts, D. (2012). Modelling withdrawal and persistence for initial teacher training: Revising Tinto's longitudinal model of departure. *British Educational Research Journal, 38*(6), 953–975. https://doi.org/10.1080/01411926.2011.603035

Rubin, D. B. (1987). *Imputation for nonresponse in surveys*. John Wiley & Sons, Inc.

Salmela-Aro, K., & Upadyaya, K. (2014). School burnout and engagement in the context of demands-resources model. *British Journal of Educational Psychology, 84*(1), 137–151. https://doi.org/10.1111/bjep.12018

Salmela-Aro, K., Tang, X., & Upadyaya, K. (2022). Study demands-resources model of student engagement and burnout. In *Handbook of Research on Student Engagement* (pp. 77–93). https://doi.org/10.1007/978-3-031-07853-8_4

Santana-Monagas, E., da Costa Ferreira, P., Veiga Simão, A. M., & Núñez, J. L. (2024). How (de)motivating teaching styles shape message framing outcomes on students' self-efficacy, emotions, and grades. *Learning and Individual Differences*. https://doi.org/10.1016/j.lindif.2024.102420

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal, 43*(6), 1248–1264.

Schilke, O., Homburg, C., Klarmann, M., & Reimann, M. (2012). What drives key informant accuracy? *Journal of Marketing Research*, *49*, 594–608.

Smith, C. D., & Baik, C. (2021). High-impact teaching practices in higher education: A best evidence review. *Studies in Higher Education*, *46*(8), 1696–1713. https://doi.org/10.1080/03075079.2019.1698539

Solberg, V. S., O'brien, K., Villareal, P., Kennel, R., & Davis, B. (1993). Self-efficacy and Hispanic college students: Validation of the college self-efficacy instrument. *Hispanic Journal of Behavioral Sciences, 15*(1), 80–95. https://doi.org/10.1177/07399863930151004

Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers and Education, 46*(1), 29–48. https://doi.org/10.1016/j.compedu.2005.04.002

Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). Knowledge discovery out of text data: A systematic review via text mining. *Journal of Knowledge Management*, *22*(7), 1471–1488. https://doi.org/10.1108/JKM-11-2017-0517

Vohs, K. D., Baumeister, R. F., & Ciarocco, N. J. (2005). Self-regulation and self-presentation: Regulatory resource depletion impairs impression management and effortful self-presentation depletes regulatory resources. *Journal of Personality and Social Psychology*, *88*(4), 632–657. https://doi.org/10.1037/0022-3514.88.4.632

Walker, D., & Myrick, F. (2006). Grounded theory: An exploration of process and procedure. *Qualitative Health Research*, *16*(4), 547–559. https://doi.org/10.1177/1049732305285972

Watt, H. M. G., & Richardson, P. W. (2007). Motivational factors influencing teaching as a career choice: Development and validation of the FIT-choice scale. *Journal of Experimental Education*, *75*(3), 167–202. https://doi.org/10.3200/JEXE.75.3.167-202

Yang, C., Sharkey, J. D., Reed, L. A., Chen, C., & Dowdy, E. (2018). Bullying victimization and student engagement in elementary, middle, and high schools: Moderating role of school climate. *School Psychology Quarterly*, *33*(1), 54–64. https://doi.org/10.1037/spq0000250

Yerkes, R.M., & Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology & Psychology, 18*, 459–482. https://doi.org/10.1002/cne.920180503

Yu, X., Cheng, H., Liu, X., Roth, D., & Gao, J. (2023). Automatic hallucination assessment for aligned large language models via transferable adversarial attacks. *arXiv.* http://arxiv.org/abs/2310.12516

Zecca, G., Györkös, C., Becker, J., Massoudi, K., De Bruin, G. P., & Rossier, J. (2015). Validation of the French Utrecht work engagement scale and its relationship with personality traits and impulsivity. *European Review of Applied Psychology, 65*(1), 19–28. https://doi.org/10.1016/j.erap.2014.10.003

Authors and Affiliations

Fernando Núñez-Regueiro[1] · Samuel Falcon[2,3] · Pascal Bressoux[1]

✉ Samuel Falcon
samuel.falcon@pdi.atlanticomedio.es

Fernando Núñez-Regueiro
fernando.nunez-regueiro@univ-grenoble-alpes.fr

Pascal Bressoux
pascal.bressoux@univ-grenoble-alpes.fr

[1] Contextual Learning Research Laboratory (LaRAC), Université Grenoble Alpes, Saint-Martin-d'Hères 38400, France

[2] Department of Psychology, University of Atlántico Medio, Ctra. de Quilmes, 37, Las Palmas de Gran Canaria 35017, Spain

[3] Instituto Universitario de Análisis y Aplicaciones Textuales (IATEXT), University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria 35003, Spain