— LIFECYCLE —
JOURNAL

Student Perspectives on High-Stakes Testing (2015–2025): An AI-Assisted Systematic Review of Empirical Evidence on Motivation, Learning, and Well-Being

Juda Kaleta 1, 2

1. Institute of History, Faculty of Arts, Charles University, Czech Republic
2. Doctorado en Educación, Universidad de Murcia, Spain

Published on 11.04.2025
Research Plan
DOI: *10.71240/lcyc.736516*

Abstract

This study proposes an AI-assisted systematic literature review of empirical research on student experiences with high-stakes testing published between 2015 and 2025. Despite ongoing debates about the "culture of testing," there remains a lack of synthesized evidence foregrounding students' perspectives, particularly regarding motivation, learning strategies, and well-being. This review aims to identify, classify, and synthesize studies that collected primary data from students (e.g., questionnaires, interviews, or mixed-methods designs) and examined the psychological and educational effects of standardized or high-stakes assessments.

The review protocol integrates washback theory, which captures the systemic and classroom-level consequences of testing, with self-determination and expectancy-value theories, which provide a framework for understanding how test contexts shape intrinsic and extrinsic motivation. The

review will focus primarily on studies from Europe, while including a limited number of non-European studies to illustrate cross-cultural variation.

The AI-assisted component will be applied to three stages: (1) abstract screening against predefined inclusion criteria; (2) structured extraction of study characteristics (country, sample, methodology, theoretical framework, outcomes); and (3) preliminary thematic clustering of findings. All AI outputs will undergo systematic human validation of a random subset to evaluate accuracy and reliability.

Data synthesis will combine descriptive mapping of studies with thematic analysis structured around washback, motivation, and well-being constructs. The review is expected to provide (a) a comprehensive overview of student experiences with high-stakes testing, (b) insight into the motivational and affective mechanisms underlying test-related behaviors, and (c) an evaluation of the feasibility and transparency of AI-assisted evidence synthesis in educational research.

Registration

N/A - This is Stage 1 Registration.

Materials

Materials are not yet available because data collection and AI-assisted extraction have not been performed. All relevant materials will be uploaded to an open repository (e.g., OSF) once the study is initiated.

Data

Materials are not yet available because data collection and AI-assisted extraction have not been performed. All relevant materials will be uploaded to an open repository (e.g., OSF) once the study is initiated.

Code

Materials are not yet available because data collection and AI-assisted extraction have not been performed. All relevant materials will be uploaded to an open repository (e.g., OSF) once the study is initiated.

Paper

Student Perspectives on High-Stakes Testing (2015–2025): An AI-Assisted Systematic Review of Empirical Evidence on Motivation, Learning, and Well-Being

Juda Kaleta

Introduction

Background and Rationale

High-stakes testing—examinations that determine academic progression, certification, or institutional accountability—has become a defining feature of contemporary education systems. From standardized national assessments to selective entrance exams, these tests are promoted as tools for fairness, efficiency, and accountability (Harlen & Deakin Crick, 2003; Koretz, 2010). Yet their societal and psychological consequences have been increasingly contested. Empirical studies link high-stakes regimes to heightened anxiety (Putwain et al., 2016), surface-level or strategic learning strategies (Muth & Lüftenegger, 2025; Popham, 2001), and widening achievement gaps (Reardon, 2018).

Policymakers continue to rely on test-based evidence to guide reforms, creating a culture of testing—an educational environment in which assessment outcomes dominate institutional priorities and student identities (Au, 2007; Bennett, 2011). Within this culture, the student perspective often remains marginal, with much research focusing on teacher behavior, curriculum narrowing, or system-level effects.

Understanding students' experiences is critical because motivation and self-regulation are central to learning outcomes (Eccles & Wigfield, 2020; Ryan & Deci, 2000). Students may simultaneously experience increased competence or expectancy for certain tasks while facing reduced autonomy or heightened anxiety, illustrating the complex effects of high-stakes testing.

A systematic synthesis of the literature is timely. The past decade (2015–2025) has seen major developments in educational policy, motivation research, and AI-assisted review methods (Grootendorst, 2022). This project will consolidate empirical evidence on student-level outcomes of high-stakes testing, focusing on motivation, learning strategies, and well-being, using an AI-assisted systematic review approach.

Theoretical Frameworks

Washback theory

The concept of washback (Alderson & Wall, 1993; Green, 2013) describes the influence that tests exert on teaching and learning. While originally framed in applied linguistics, the construct has been extended to general education to conceptualize how assessment systems shape both instruction and learner behavior. Washback may be positive, stimulating productive learning behaviors aligned with curricular goals, or negative, inducing rote learning, anxiety, and strategic compliance.

Most washback research has examined teachers' instructional adjustments, leaving the student dimension comparatively underexplored. When students are considered, findings show wide variation: some report loss of intrinsic motivation and autonomy, others describe heightened effort and achievement orientation. These inconsistencies suggest that washback interacts with deeper motivational processes—offering a bridge to psychological theories such as SDT and EVT.

Self-Determination Theory (SDT) & Expectancy-Value Theory (EVT)

*Self-Determination Theory (SDT) (*Ryan & Deci, 2000) posits that motivation exists along a continuum from intrinsic to extrinsic, depending on the degree to which behavior is self-endorsed. Three basic psychological needs—autonomy, competence, and relatedness—govern this continuum. When high-stakes testing constrains autonomy or threatens competence, students may shift from autonomous to controlled motivation, resulting in performance anxiety or disengagement.

*Expectancy-Value Theory (EVT) (*Eccles & Wigfield, 2020) complements this view by emphasizing students' beliefs about success (expectancy) and the perceived value of the task (value). High-stakes exams can simultaneously raise value (due to importance or consequences) while lowering expectancy (due to difficulty or competition). The resulting motivational patterns—ranging from defensive effort to withdrawal—may help explain heterogeneous washback effects across contexts.

Integrating SDT and EVT provides a dual lens: SDT clarifies why certain motivational states arise under external pressure, while EVT explains how students weigh effort and outcome expectations. This synthesis provides a more nuanced understanding of the effects of testing on learning strategies and well-being.

Recent Conceptual Developments

A recent review by Muth & Lüftenegger (2025) argues that debates around testing have become increasingly ideological, with claims about "teaching to the test" often outpacing empirical

evidence. Their analysis highlights two critical gaps: (a) a shortage of high-quality, student-centered data, and (b) limited integration of motivational theory into empirical assessment research. The authors call for systematic evidence that disentangles contextual, cultural, and psychological mechanisms underpinning testing effects.

This Stage 1 proposal directly responds to that call. By systematically synthesizing a decade of empirical findings through both human and AI-assisted analysis, the study aims to clarify the magnitude and variability of motivational consequences associated with high-stakes assessment worldwide.

Motivation for AI-assisted methods in systematic reviews

Traditional systematic reviews in education rely on manual screening and coding, which are labor-intensive and prone to selection bias. Advances in natural language processing (NLP) and large language models (LLMs) now enable scalable extraction and clustering of research evidence. AI-assisted methods can identify latent thematic structures, suggest inclusion decisions, and generate preliminary codes with human oversight.

In the present study, AI is not a replacement for human judgment but an augmentation of it. The workflow will use LLM-based semantic search and clustering (BERTopic) to pre-classify abstracts and full texts. Human reviewers will then verify inclusion, refine themes, and assess accuracy through inter-rater reliability metrics (Cohen's κ). This dual process allows both substantive insights and a methodological test of whether AI-assisted synthesis can meet transparency and reproducibility standards expected in educational psychology.

Research Objectives and Scope

The project has *two main objectives:*

1. *Substantive Objective:*

   To synthesize empirical findings from 2015 to 2025 that investigate how high-stakes testing affects students' *motivation, learning strategies,* and *well-being.* The review will identify dominant themes, theoretical framings, and regional variations, and emphasize how SDT and EVT constructs manifest across different washback contexts.

2. *Methodological Objective:*

To evaluate the reliability and efficiency of *AI-assisted evidence extraction* compared with traditional human coding. Agreement rates (Cohen's $\kappa \geq 0.85$) will be used as a criterion for methodological adequacy.

Scope and Inclusion Criteria:

- Empirical peer-reviewed studies (quantitative, qualitative, or mixed methods).
- Published between *2015–2025* in English.
- Focus on *students* (K–12 or tertiary) as primary data sources (e.g., surveys, interviews, think-alouds).
- Address high-stakes or standardized testing contexts.

Studies focusing exclusively on teachers, institutions, or policy-level outcomes without student data will be excluded.

By situating student experiences within the broader washback framework and grounding the synthesis in SDT and EVT, this review will contribute both theoretical clarity and methodological innovation. The ultimate goal is to move beyond polarized narratives about "teaching to the test" toward a more empirically grounded understanding of how assessment regimes shape the motivational lives of learners.

Research Questions

The present study addresses three interrelated research questions that jointly examine the psychological and methodological dimensions of student experiences with high-stakes testing.

- *RQ1. What patterns and themes emerge in empirical studies (2015–2025) that investigate students' motivation, learning, or well-being in relation to high-stakes testing?*

  This question aims to map the global empirical evidence on how students perceive and experience testing regimes. Using AI-assisted topic modeling (BERTopic), the review will identify dominant thematic clusters, theoretical frameworks, and methodological trends across regions and education levels.

- *RQ2. How do motivational mechanisms derived from Self-Determination Theory (SDT) and Expectancy-Value Theory (EVT) manifest in students' reported experiences of high-stakes testing?*

  Building on the washback framework, this question explores whether testing contexts predominantly undermine or support autonomy, expectancy, and task value. The analysis

will quantify the proportion of studies reporting positive, negative, or mixed motivational outcomes, allowing comparison across cultural and institutional contexts.

- *RQ3. How reliable and efficient are AI-assisted text-analysis methods compared with human coding in systematic reviews of educational psychology research?*

  This question assesses the methodological contribution of AI-assisted synthesis. Agreement between AI and human coders (Cohen's κ) will determine whether automated extraction can meet standards of accuracy and transparency suitable for preregistered evidence synthesis.

Together, these questions integrate three complementary goals:

1. *Descriptive:* Characterize the empirical knowledge base on student experiences with high-stakes testing (RQ1).
2. *Explanatory:* Evaluate motivational mechanisms through the lens of SDT and EVT (RQ2).
3. *Methodological:* Test the reliability and efficiency of AI-assisted review workflows (RQ3).

Addressing these questions will provide both substantive insight into how testing affects students' motivational lives and procedural guidance for how emerging AI tools can support transparent, scalable evidence synthesis in educational psychology.

Methods

Study Design Overview

This study is a preregistered, systematic literature review (SLR) of empirical research published between 2015 and 2025 that examines the effects of high-stakes testing on student motivation, learning, and well-being. The review integrates washback theory with Self-Determination Theory (SDT) and Expectancy-Value Theory (EVT) as conceptual lenses. An AI-assisted screening and extraction workflow (BERTopic-based topic modeling) will be applied to streamline literature processing while ensuring reproducibility and transparency.

The design follows a Stage 1 Registered Report framework, pre-specifying research questions, inclusion criteria, sampling procedures, analysis plans, stopping rules, and reliability thresholds for AI-assisted coding.

Eligibility Criteria

Inclusion Criteria

- Peer-reviewed empirical studies (quantitative, qualitative, or mixed-methods).

- Published 2015–2025 in English.
- Primary focus on students (K–12 or tertiary education).
- Examine high-stakes or standardized testing contexts, including national exams, entrance tests, or high-consequence classroom assessments.
- Report on at least one of: motivation, learning strategies, or well-being.

Exclusion Criteria

- Studies reporting only teacher, policy, or system-level outcomes without student-level data.
- Non-empirical publications (reviews, conceptual papers, commentaries).
- Non-English publications or pre–2015 publications.

Search Strategy

The search will be conducted in *Scopus* and *ERIC,* using the following query:

(“high-stakes test*” OR “standardized assessment*” OR “exam pressure”) AND (student* AND (motivation OR learning OR “well-being” OR engagement)) AND PUBYEAR > 2014 AND PUBYEAR < 2026

- The title, abstract, and keywords fields will be searched.
- Search results will be exported with metadata (title, authors, year, DOI, abstract, keywords).
- Duplicates will be removed automatically.
- After title/abstract screening, an eligible pool of studies will be identified.
- Random sampling with a specified seed will select N = 100 full-text studies for analysis, ensuring regional stratification with European studies at 50% of the sample.

AI-Assisted Screening and Extraction

- *Screening:* All abstracts and titles will be processed using BERTopic for semantic clustering.
    - Inclusion suggestions will be generated automatically, followed by manual validation to confirm eligibility.

- *Full-text extraction:* BERTopic will identify recurrent themes; representative documents will be manually checked.
- *Motivational coding:* SDT (autonomy, competence, relatedness) and EVT (expectancy, value) constructs will be coded as ↑ / ↓ / neutral / mixed.
- *Validation subset:* 25 randomly selected studies (25% of N = 100) will be independently coded by two human coders to compute Cohen's κ. Thresholds: κ ≥ 0.85 (acceptable), 0.70–0.85 (partial reliability, extend validation), <0.70 (revert to human-only coding).

Data Synthesis

- *Descriptive synthesis:* Count of studies by year, region, method, and outcome type.
- *Thematic synthesis:* AI-generated clusters reviewed and refined by humans; themes occurring in ≥10 studies considered robust.
- *Motivational patterns:* Frequencies of SDT/EVT constructs aggregated and cross-tabulated by region and test type.
- *Interpretive framework:* Integration of washback theory with SDT/EVT to explain observed patterns of student motivation, learning, and well-being.

Risk of Bias & Quality Assessment

To ensure the credibility, transparency, and reproducibility of the systematic review, each included study will be assessed for risk of bias and methodological quality using a structured, multi-step procedure adapted from the Mixed Methods Appraisal Tool (Hong et al., 2018) and best practices in educational research synthesis.

Domains of Risk Assessment

1. **Sampling and Participant Selection**
    - Criterion: Whether the study population is representative of the target student group.
    - Assessment: Check for clarity of inclusion/exclusion criteria, recruitment methods, sample size justification, and demographic reporting.
    - Risk Levels:
        - Low: Random or stratified sampling, clear inclusion criteria.

- Moderate: Convenience sampling with reasonable justification.
- High: Unclear or biased selection limiting generalizability.

2. **Measurement and Instrumentation**
   - Criterion: Validity and reliability of instruments used to measure motivation, learning, or well-being.
   - Assessment: Confirm whether standardized, validated questionnaires/scales or rigorously piloted interview protocols were used; check reporting of psychometric properties.
   - Risk Levels:
     - Low: Validated instruments with reported reliability.
     - Moderate: Custom instruments with some validation.
     - High: Poorly described instruments or untested measures.
3. **Data Analysis and Reporting**
   - Criterion: Appropriateness and transparency of analytical methods.
   - Assessment: Review whether the analysis aligns with the research questions, whether statistical assumptions are addressed, and whether qualitative coding procedures are explained (e.g., inter-rater agreement, thematic saturation).
   - Risk Levels:
     - Low: Clear, rigorous, and transparent methods.
     - Moderate: Partial reporting or minor inconsistencies.
     - High: Analytical procedures are insufficiently described or inappropriate.
4. **Outcome Reporting and Selective Reporting**
   - Criterion: Completeness of reported results relative to stated research questions and hypotheses.
   - Assessment: Identify missing outcomes, inconsistent reporting, or evidence of selective presentation of results.
   - Risk Levels:

- Low: All outcomes reported as planned.
- Moderate: Minor omissions or ambiguities.
- High: Major omissions affecting interpretability.

5. **Study-Level Bias Synthesis**
    - Each study will receive a composite risk-of-bias rating (Low / Moderate / High) based on the domains above.
    - Domains will be weighted equally, but critical flaws in sampling or measurement will automatically flag a study as at High risk, regardless of other domains.

Implementation Procedure

1. Independent Assessment: Two reviewers will independently evaluate each study. Disagreements will be resolved through discussion or, if needed, by a third reviewer.
2. Documentation: A structured Excel or CSV log will capture scores for each domain, comments, and final composite rating.
3. Sensitivity Analyses:
    1. Primary thematic synthesis will include all studies.
    2. Secondary analyses will exclude high-risk studies to assess the robustness of conclusions.
4. AI-Assisted Extraction Verification: Although initial thematic coding is AI-assisted, all risk-of-bias ratings are performed by humans, ensuring that automation does not compromise quality assessment.

Study Design Table

Question	**Hypothesis**	**Sampling plan**	**Analysis Plan**	**Rationale for deciding the sensitivity of the test for confirming or**		

				ng the hypothesis	**Interpretation given different outcomes**	**Theory that could be shown wrong by the outcomes**
RQ1. What patterns and themes emerge in empirical studies (2015–2025) that examine students' motivation, learning, or well-being in relation to high-stakes testing?	- (exploratory)	*Sampling frame:* Scopus + ERIC searches (English-language, 2015–2025) using terms *("high-stakes test"* OR "standardized assessment*" OR "exam pressure") AND (student* AND (motivation OR learning OR well-being))*. *Expected yield:* 800 records. *Stage 1 screening:* Title/abstract filter to include only studies with (a) student participants and (b) empirical data. *Target = 200 eligible abstracts. Final inclusion:* Randomly sample *N* =				

		100 full-text studies from the eligible set using reproducible randomization. *Inclusion strata:* Ensure representation of European studies by region to 50%.	Descriptive mapping of publication year, region, methods, and frameworks. AI-assisted thematic extraction (BERTopic) to identify recurrent conceptual clusters; human review to confirm or merge final themes. Themes with <3 mentions across studies are treated as idiosyncratic.	Sensitivity: A theme is "robust" if it appears in ≥10 independent studies (10% threshold). Thematic saturation expected around N≈100 based on prior SLRs.	Dominant negative/well-being themes → consistent evidence of adverse washback. Mixed pattern → context-contingent washback. Predominantly positive → challenge to current critical discourse.	The generalized claim that high-stakes testing uniformly undermines student learning and motivation could be shown to be incomplete if mixed or positive evidence predominates.
RQ2. How do motivational mechanisms from SDT and EVT manifest in students' reported experiences of high-stakes testing?	High-stakes contexts are associated with decreased autonomy and intrinsic motivation (SDT) and reduced expectancy/value (EVT).	Subset of M = 50 studies from the N = 100 corpus that explicitly report motivational constructs. If >50, use random sampling.	Frequency and co-occurrence analysis of coded constructs (↑ / ↓ / neutral / mixed). Cross-tabulation by region, level, and test type.	≥60% (≥30/50) reporting decreased autonomy/expectancy/value → supports hypothesis. 40–60% → mixed. ≤40% → disconfirming.	Same as above.	Strong-form SDT claims that external evaluation reliably undermines autonomous motivation.

	theme categorization.	Validation subset: 25 studies (25%) randomly drawn from the corpus. Both AI and two independent human coders assess inclusion and thematic coding.	Compute Cohen's κ, precision, recall, F1, and mean person-hours per document.	$\kappa \geq 0.85 \rightarrow$ reliable; 0.70–0.85 $\rightarrow$ partial; <0.70 $\rightarrow$ unreliable.	$\kappa \geq 0.85 \rightarrow$ AI acceptable for screening. <0.70 $\rightarrow$ revert to human-only.	That LLMs can autonomously conduct reliable SLR coding without supervision.

Analysis Plan

The analysis plan is structured around three interconnected dimensions: quantitative synthesis, qualitative thematic analysis, and AI-method evaluation. Each is described below with explicit metrics, thresholds, and procedures.

Quantitative Synthesis

- Objective: Summarize study characteristics and outcomes numerically to identify patterns across regions, test types, and student populations.
- Procedures:
    1. Extract metadata from all included studies: publication year, country/region, education level, sample size, study design (quantitative/qualitative/mixed-methods), and outcome measures.
    2. Compute frequency distributions for categorical variables (e.g., region, test type, outcome category).
    3. Generate descriptive statistics (mean, median, range) for continuous variables (e.g., sample size, scale scores).
    4. Perform regional comparisons using contingency tables and chi-square tests (or Fisher's exact test if expected cell counts <5) to explore differences in reported motivational or well-being outcomes across regions (Europe, Asia, Other).

- Interpretation thresholds:
    - Dominant trends are identified if ≥60% of studies in a category report similar directional outcomes (e.g., decreased autonomy, increased effort).
    - Mixed or inconsistent patterns indicate if directional outcomes are roughly balanced (40–60%).

Qualitative Thematic Synthesis

- Objective: Identify recurring patterns in students' reported experiences with high-stakes testing, focusing on motivation, learning strategies, and well-being.
- Procedures:
    1. Apply BERTopic to full-text documents to generate initial clusters of semantically similar content.
    2. Human reviewers validate and refine clusters to create coherent themes.
    3. Code each study for SDT constructs (autonomy, competence, relatedness) and EVT constructs (expectancy, value) as:
        1. ↑ (enhanced), ↓ (diminished), 0 (neutral/no effect), or X (mixed)
    4. Aggregate counts across studies and cross-tabulate by region, test type, and educational level.
    5. Conduct thematic saturation check: additional studies included until no new themes emerge in two consecutive randomly sampled studies.
- Interpretation thresholds:
    - A theme is considered robust if it appears in ≥10 studies (≥10% of N = 100).
    - Less frequent patterns (<10 studies) are treated as context-specific or exploratory.

AI-Assisted Method Evaluation

- Objective: Assess the reliability, efficiency, and accuracy of the AI-assisted extraction workflow relative to human coding.
- Procedures:

1. Select 25 studies (25% of N = 100) randomly for dual coding by AI and two human coders.
2. Compute Cohen's κ for inclusion/exclusion decisions and theme assignment. Thresholds:
3. $\kappa \geq 0.85 \rightarrow$ high reliability, AI output accepted.
4. $0.70 \leq \kappa < 0.85 \rightarrow$ moderate reliability, expand validation set to 50 studies.
5. $\kappa < 0.70 \rightarrow$ insufficient reliability, revert to human-only coding.
6. Measure efficiency gains: average human hours per study with AI assistance vs. manual-only coding.
7. Report accuracy metrics—precision, recall, and F1 score—for inclusion and thematic classification.

- Stopping Rules & Rationale:
  - If AI fails to reach $\kappa \geq 0.70$ in the validation subset, the workflow will be discontinued to preserve data integrity.
  - If thematic saturation is reached before N = 100, the synthesis continues to the full sample to ensure complete coverage of SDT/EVT constructs.

Expected Outcomes

1. *Overview of Student Experiences*

   A comprehensive map of empirical findings from 2015–2025 on how students perceive and experience high-stakes testing, including differences across regions, educational levels, and test types.

2. **Motivation, Learning, and Well-Being Insights**
   - Identification of patterns in SDT (autonomy, competence, relatedness) and EVT (expectancy, value) constructs.
   - Quantification of positive, negative, neutral, and mixed effects on student motivation.
   - Characterization of associated outcomes in learning strategies, engagement, and psychological well-being, highlighting context-dependent washback effects.

3. **AI-Assisted SLR Feasibility and Transparency**
    - Evaluation of the accuracy, reliability, and efficiency of AI-assisted extraction (BERTopic + human validation) relative to manual coding.
    - Assessment of whether AI assistance supports reproducible, transparent systematic reviews without compromising data quality.

Together, these outcomes will provide theoretical clarity, empirical synthesis, and methodological guidance for understanding student-level consequences of high-stakes testing and the utility of AI-assisted evidence synthesis in educational psychology.

Ethics Statement

No ethics approval was required or obtained.

Generative AI (ChatGPT, Grammarly) supported brainstorming, language editing, and stylistic refinement. All research design, intellectual contributions, and interpretations are solely our own.

Acknowledgments

This work was supported by the project “Human-centred AI for a Sustainable and Adaptive Society” (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union.

Fundings

The funder(s) played no role in the study design, data collection, analysis, publication decision, or submission preparation.

Conflict of interest

The author declares that they have no conflicts of interest.

References

Alderson, J. C., & Wall, D. (1993). Does Washback Exist? *Applied Linguistics, 14(*2), 115–129. *https://doi.org/10.1093/applin/14.2.115*

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, 36(*5), 258–267. *https://doi.org/10.3102/0013189x07306523*

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18(*1), 5–25. *https://doi.org/10.1080/0969594X.2010.513678*

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61,* 101859. *https://doi.org/10.1016/j.cedpsych.2020.101859*

Green, A. (2013). Washback in language assessment. *International Journal of English Studies, 13(*2). *https://doi.org/10.6018/ijes.13.2.185891*

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (Version 1). arXiv. *https://doi.org/10.48550/ARXIV.2203.05794*

Harlen, W., & Deakin Crick, R. (2003). Testing and Motivation for Learning. *Assessment in Education: Principles, Policy & Practice, 10(*2), 169–207. *https://doi.org/10.1080/0969594032000121270*

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34(*4), 285–291. *https://doi.org/10.3233/EFI–180221*

Koretz, D. (Ed.). (2010). *Measuring up: What educational testing really tells us.* Harvard University Press.

Muth, J., & Lüftenegger, M. (2025). Teaching to the test: Unraveling the consequences for student motivation. *Learning and Individual Differences, 121,* 102707. *https://doi.org/10.1016/j.lindif.2025.102707*

Popham, W. J. (2001). *The Truth about Testing: An Educator's Call to Action.* Association for Supervision & Curriculum Development.

Putwain, D. W., Symes, W., & Remedios, R. (2016). The impact of fear appeals on subjective-task value and academic self-efficacy: The role of appraisal. *Learning and Individual Differences, 51,* 307–313. *https://doi.org/10.1016/j.lindif.2016.08.042*

Reardon, S. F. (2018). The Widening Academic Achievement Gap Between the Rich and the Poor. In D. B. Grusky & J. Hill (Eds), *Inequality in the 21st Century* (1st edn, pp. 177–189). Routledge. *https://doi.org/10.4324/9780429499821–33*

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55(*1), 68–78. *https://doi.org/10.1037/0003–066X.55.1.68*

Declarations

Ethics

No ethics approval was required or obtained.

Generative AI (ChatGPT, Grammarly) supported brainstorming, language editing, and stylistic refinement. All research design, intellectual contributions, and interpretations are solely our own.

Competing Interests

The author declares that they have no conflicts of interest.

Funding

This work was supported by the project “Human-centred AI for a Sustainable and Adaptive Society” (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union. The funder(s) played no role in the study design, data collection, analysis, publication decision, or submission preparation.

Author Contributions

Juda Kaleta conceived and designed the research question, defined the inclusion/exclusion criteria, and developed the systematic review protocol including AI-assisted screening and extraction procedures. He prepared the planned data extraction forms, coding schemes, and thematic synthesis framework aligned with washback theory and SDT/EVT. He drafted the Stage 1 registration document and will be responsible for conducting the literature search, AI-assisted screening, data extraction, coding, and synthesis once the review is approved. CRediT roles: Conceptualization, Methodology, Software (AI-assisted extraction design), Project administration, Writing – original draft (registration), Writing – review & editing.