Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The depth of washback in Vietnamese EFL teacher responses to the reformed national high school graduation English exam before its first administration

Khanh Nguyen-Xuan

khanhnx@thanhdong.edu.vn

Thanh Dong University

Thuyen Do Ngoc-Nhu

ACE The Future JSC

Thang Nguyen Van

Thanh Dong University

Research Article

Keywords: technical washback, depth of engagement, high-stakes testing, teacher response, EFL, Vietnam

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11334749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Washback research has described the direction, intensity and variability of test effects on teaching, but it has said less about depth, that is, how far teachers' responses engage with the construct a test sets out to measure. In this article, we argue that depth deserves attention as a distinct dimension of washback. We develop the argument through the 2025 redesign of the English paper in Vietnam's National High School Graduation Exam (NHGE), which replaced discrete grammar and vocabulary items with tasks set in whole texts and gave teachers about eight months to prepare. Using an exploratory qualitative comparative case design, we interviewed seven Grade 12 EFL teachers in urban and rural settings before any results existed, through focus groups and follow-up interviews. We analysed the data through reflexive thematic analysis and read them alongside the exam papers. In both settings, teachers invested most of their effort in copying item formats and much less in the discourse competence those formats sample. We propose the term technical washback for this pattern of effortful but surface-level compliance. Urban teachers turned compliance into dual-track experimentation, whereas rural teachers described a retreat toward grammar translation, and many rural students chose not to sit English once it became optional. Building on these findings, we offer a working framework that crosses intensity with depth. We conclude that assessment reform needs professional development that equips teachers with pedagogical techniques for teaching the construct, not only skills in writing test items.

1 Introduction

High-stakes examinations can shape classroom practice in ways that curricula alone do not. Reviews of washback research have documented this influence across many systems, and they also show that it is rarely simple or predictable [1, 2]. Recent studies from quite different contexts illustrate the point. In Iran, researchers have linked the university entrance examination to schooling narrowed toward test preparation and memorisation [3], and they have associated a receptive-only English proficiency test with rote learning [4]. In Ethiopia, a newly introduced exit examination led university instructors to reorganise their teaching around exam content [5]. In each of these cases, the test itself formed part of the problem. We are interested in a different situation, in which a ministry redesigns a test so that it measures what the curriculum values. If the new test is better, will teaching improve with it?

We consider the English paper of Vietnam's National High School Graduation Exam (NHGE) a timely case for this question. On 18 October 2024, the Ministry of Education and Training (MOET) released a revised reference exam that removed every discrete grammar, vocabulary and pronunciation item [6]. In their place, the paper introduced cloze tasks in real-life genres, ordering tasks built from conversations, letters and paragraphs, and inference, paraphrase and summary items. The official English paper of June 2025 confirmed this design [7]. At the same time, the paper shrank from 50 to 40 items, and English changed from a compulsory to an optional exam subject [8]. Teachers therefore had about eight months to prepare students for a more demanding test that students could now choose not to take. The reform also belongs to a national agenda that aims to make English a second language in schools between 2025 and 2035 [9, 10].

The redesign responded to a long-standing problem. For over a decade, discrete-point items had dominated the English paper [11], even after the 2018 General Education Curriculum made communicative competence the goal of school English [12, 13]. In our reading, the 2025 paper is the first to align with that goal. However, a better test reaches classrooms only through teachers, who respond with the ideas, resources and time they have [14]. Across Asia, reforms aimed at communicative competence have repeatedly met a gap between policy design and what teachers feel prepared to do [15, 16]. Vietnamese research has shown a similar gap in curriculum reform, where teachers' agency depended on school support, workload and their own understanding of the new pedagogy [17–19]. Surprisingly, however, researchers have paid little attention to the 2025 English paper. To our knowledge, the only published study of its washback surveyed students after the first administration [20]. We found no study of how teachers responded while the reform was unfolding, and none that compared urban and rural settings, although English provision in Vietnam differs sharply between the two [21].

In this study, we address these gaps. We examine how Grade 12 EFL teachers in urban and rural settings responded to the redesigned paper in the months before its first administration, and we propose concepts that later studies can test. Two research questions guided our work.

RQ1: How did EFL teachers in urban and rural Vietnam perceive the 2025 NHGE English reform during its implementation phase?

RQ2: How did they respond, and what does the depth of that response reveal about the relationship between format change and pedagogical change?

2 Literature review

2.1 Washback theory and the question of depth

Early washback research began with a cautionary point. Messick [22] argued that tests which look authentic do not guarantee productive teaching, and Alderson and Wall [1] showed that washback is neither automatic nor uniform. Alderson and Hamp-Lyons [23] then observed the same teachers in test-preparation and regular classes, and they concluded that simple hypotheses about washback could not explain what they saw. Building on this work, Bailey [24] proposed a basic model in which a test acts on four groups of participants, namely students, teachers, materials writers and curriculum designers, and researchers. Their processes lead to products such as learning, teaching, new materials and research results, which in turn feed back to the test. We use this model as a sensitising lens, because it helps us see where a nationwide reform has arrived and where it has not.

Later work described the qualities of washback in more detail. Green [25] distinguished three of them. Direction depends on the overlap between the test and the focal construct, that is, the target skills as course providers and learners understand them. Intensity refers to the degree to which participants adjust to test demands, and it rises when a test is important, challenging but attainable, and close in time. Variability recognises that washback "is not deterministic" [25, p. 18], since participants differ in

knowledge, resources and acceptance of test demands. Other models place the teacher at the centre. Shih [26] modelled how school, teacher and test factors together shape whether and how teaching changes, and Spratt [27] identified the teacher as a key mediating factor. More recently, Cheng et al. [2] reviewed 87 empirical studies and placed washback within an argument-based approach to validity, in which the consequences of test use become evidence about a test.

In our view, these models explain well why washback varies, but they say less about what varies within a single teacher's response. Throughout this paper, we use the construct to mean the ability a test sets out to measure, which for the 2025 English paper is mainly the ability to read and use connected text. Green's direction concerns how far a test overlaps with this construct, so it describes the test and its use. Intensity concerns how much teachers adjust. Neither dimension tells us whether a teacher's adjustment engages with the construct itself. Interestingly, several studies have reported cases in which it did not. Wall [28] found that Sri Lankan teachers adopted the formats of a new exam while keeping their old methods, and Qi [29] showed that conflicting stakeholder aims weakened the intended washback of a Chinese national test. Outside language testing, Coburn [30] argued that four dimensions define the scale of a reform, namely depth, sustainability, spread and a shift in ownership of the reform. Of these, depth refers to change in teachers' beliefs and pedagogical principles, not only in materials and activities. Taken together, these studies lead us to suspect that a test may overlap well with its construct and still produce shallow change, particularly when teachers have had no preparation for teaching that construct. We use the term depth for this missing link between the test's construct and the teacher's practice.

We also consider timing an important and often overlooked factor. Most washback studies look back after results appear, and as a result they rarely capture how teachers make sense of a new construct before outcomes exist. However, this is precisely the period in which teachers build materials, schools rewrite their tests and education authorities make decisions about professional development.

2.2 The reformed English paper and its construct

In models of communicative competence, discourse competence is the ability to combine forms and meanings into unified texts. Its components include cohesion, deixis, coherence and generic structure [31, 32]. Our comparison of the papers (ESM, Table S1) shows that every item in the reference exam and the 2025 paper sits inside a text, whereas only 17 of 50 items did so in the 2024 paper. The new paper does not test discourse competence alone. Its cloze items still test grammar and vocabulary, although now within authentic genres, and its reading passages test comprehension and inference. Nevertheless, we see discourse competence as the thread that connects the tasks. Ordering tasks test cohesion and generic structure, and paraphrase and summary items test whether candidates can keep meaning constant across wordings. The paper remains indirect and receptive, with no listening, speaking or extended writing, and critics elsewhere have argued that receptive-only designs pull teaching toward test preparation [4]. Even so, teaching toward the new paper asks teachers to make cohesion, coherence and genre visible to learners, and this demand assumes that teachers themselves command these concepts.

The policy setting added a further challenge. National Assembly Resolution 88/2014/QH13 allowed several approved textbook series under one curriculum [33], so schools in the 2024 to 2025 school year taught from different books. The reference exam adapted texts from several of these series, among them Global Success, i-Learn Smart World, Friends Global and Bright [6]. In December 2025, MOET selected one series for nationwide use from the 2026 to 2027 school year [34]. The teachers in our study, however, prepared the first reformed cohort while materials still varied.

2.3 Teacher agency under reform

Whether teachers can meet a new construct is also, in our view, a question of agency. In the ecological view, agency emerges from the interplay of personal capacity and the setting [35]. It works through practical-evaluative judgement, that is, judgement about what is possible here and now [36]. Vietnamese studies support this view. Teachers enact reform selectively in light of workload and local expectations [18], their agency under the multiple-textbook policy varied with school support [17], and hierarchy and collegial networks shape how they act [37]. Teachers also live reform emotionally. Vu and Nguyen [19] found that negative emotions and fragmented understanding of communicative pedagogy led teachers to comply only superficially with a curriculum reform. We therefore expected an assessment reform to pass through unequal settings of support, and in Vietnam these settings differ markedly between cities and provinces.

3 Methodology

3.1 Research design and researcher positionality

We adopted an exploratory qualitative comparative case design, following Stake's [38] view of a case as a bounded system. Each case is a group of Grade 12 EFL teachers working under the same reform in the 2024 to 2025 school year, bounded by setting. The urban case covers teachers in the centrally governed cities, and the rural case covers teachers in provinces outside those cities. Within these boundaries, we used reflexive thematic analysis, which Braun and Clarke [39] describe as flexible enough to work within different designs. Because we aimed to capture teachers' sense-making before any results existed, we collected data in March and April 2025, after the definitive reference exam and before the June exam.

Our interest in the study grew from our own teaching. Two of us are university EFL lecturers, and one of us teaches university and high school students independently. All three of us sat the NHGE as learners. In our university classes, we have repeatedly met students who entered with high graduation scores in English but still struggled to communicate. This experience led us to ask whether the redesigned paper would change what happens in Grade 12 classrooms. Our closeness to the participants' work gave us access and a shared vocabulary, which reflexive thematic analysis treats as a resource [39]. To guard against over-identifying with teachers, the first author wrote a reflexive memo after each interview and each discussion among the authors. We also challenged one another's readings and returned to the exam papers as an external point of reference.

3.2 Participants and data collection

We reached participants through acquaintances of the research team, who introduced us to Grade 12 EFL teachers they knew. These acquaintances did not take part themselves, and none of the participants had worked with us before. We invited twelve teachers and offered no incentives. Seven took part, and the other five withdrew because of examination-season workload. We sampled purposively along one axis, urban versus rural, because English provision in Vietnam is unequal in exposure, tutoring markets, teaching time and access to international test preparation [21]. We judged the group well suited to the study's narrow aim, since every participant was preparing the first reformed cohort, which gives a sample of this kind high information power [40]. To protect participants, we report only their setting and years of teaching (Table 1).

Table 1
Participant profile

Pseudonym	Setting	Years teaching	Interviews
Grace	Urban, north	17	Group and follow-up
Lily	Urban, north	20	Group
Nancy	Urban, south	16	Group
Jessie	Urban, south	22	Group
Alice	Rural, south	10	Group and follow-up
Everlyn	Rural, south	4	Group
Rose	Rural, north	10	Group

We drew on three sources of data. First, we held one focus group interview per case in March 2025 to capture collective sense-making. Focus groups suited the study because Vietnamese teachers in a subject group plan lessons and write tests together, so group talk mirrors the unit in which adaptation happens. Second, we conducted individual follow-up interviews of about 30 minutes to enrich the group data [41]. We invited all seven teachers to a follow-up, and Grace and Alice were able to arrange one in late April 2025. We held the follow-ups before the exam because we wanted to capture responses that teachers formed without knowledge of results. A few weeks after the group interviews, teachers could also tell us how their practice had developed. Third, we used the exam papers to ground our analysis of the construct. All three of us coded each item in the 2024 paper, the October 2024 reference exam and the 2025 paper for task type, whether it stood alone or sat in a text, and the competence it most directly tested (ESM, Table S1).

The first author moderated both focus groups and Grace's follow-up, and the second author conducted Alice's follow-up. All interviews followed a semi-structured guide (ESM, Part A), and we held them in Vietnamese on Zoom. Before recording began, participants turned off their cameras and used

pseudonyms. We recorded the interviews with consent, transcribed them verbatim and stored the files in a Google Drive folder that only the research team could access. In reporting, we quote only what participants said in their own words, not phrases they agreed with when prompted.

3.3 Data analysis

We analysed the data through reflexive thematic analysis, in which researchers actively construct themes as patterns of shared meaning [39]. The first author coded the Vietnamese transcripts line by line in a shared spreadsheet, recording summaries, reflexive notes, in vivo codes and interpretive codes for each stretch of talk. Coding was inductive, and it attended both to reported practices and to the assumptions beneath them, such as the belief that matching the exam matrix is what preparing students means. We brought in Bailey's model [24] and the discourse competence construct only when we reviewed and defined themes. The second and third authors acted as critical friends [42] throughout, reading the codes and offering alternative readings. We kept a pattern only if it appeared in both cases or if its absence in one case was itself meaningful. The first author translated the quotations, and the co-authors checked each one against the original. We kept teachers' own professional terms, so ma trận appears as matrix, meaning the exam specification that lists task types and item counts.

4 Findings

4.1 Theme 1: Copying the template, yet missing adequate understanding of the construct

In both cases, teachers responded to the reform at the level of the item format more than the construct. When we asked Alice, a rural teacher, whether she had engaged with the concepts behind the reform, she answered directly.

I paid no attention to any of that. I just built materials following the format of 18 October, the newest format. If the sample gapped a reduced relative clause, I gapped a reduced relative clause, and if it gapped a phrasal verb, I gapped a phrasal verb. I did everything by copying the format. (Alice, group interview)

Other teachers described the same practice. Rose reported that "we all follow the matrix, and then we look for similar item types that stick to the matrix as closely as possible" (group interview). Everlyn described the mechanism in the field's own vocabulary.

Teachers are easily affected by washback effects. We teach the input based on the output. So at my school we also just stop at following the matrix, and the matrix is only Reading, so we focus only on Reading. (Everlyn, group interview)

In our reading, the training teachers received reinforced this orientation. Six of the seven teachers mentioned training for the reform, and the five who described its content said that it centred on the exam matrix and item writing. Everlyn heard from colleagues that provincial sessions dealt with "how to write

items, the exam matrix and item design, much more than how to teach". Notably, the urban case received the same signal. Nancy "wondered why a training session said little about teaching methods but drilled into writing questions in the exam format" (group interview), and Grace said that teaching methods came up "but not very deeply" (individual interview).

Urban teachers showed a more sophisticated version of the same pattern. Grace mapped each new item type onto a familiar test, observing that the ordering task "appears in the Aptis test format". However, she reduced the main cloze task to familiar terms, saying "its aim is really just to test grammar points on reading materials" (group interview). Jessie came closest to naming the construct when she argued that the gapped-clause items quietly assess writing, because "a student who knows how to develop a paragraph will, in reverse, order the text well when reading" (group interview).

We identified two conditions that kept the pattern in place. The first was a mismatch between materials and exam. Practice tests multiplied during the year, but teaching materials did not change in the same way. Alice found her school's textbook series "not designed like the exam content, so it is like teaching one way and testing another" (individual interview). The second condition was uncertainty about whether the real paper would match the sample.

We revise by feeling our way. If the sample gaps a reduced relative clause and we all teach reduced relative clauses, but the real exam uses an ordinary relative clause, it will be very hard. (Alice, individual interview)

Under these conditions, copying the format looked to teachers like a reasonable gamble, and the enacted curriculum narrowed. Everlyn's department "rearranged the topic list to remove what is not in the new format", and the rural schools converted their mid-term and final tests to the new format. Alice's account of the ordering task captures the result. "I have never been trained in any method for teaching it, so I just teach tips, whatever I know, to steer students toward the right answer" (individual interview).

4.2 Theme 2: Same effort to reach a shared aim, but different approaches between groups

Teachers in both cases worked hard, but their settings shaped what that effort could become. In the urban case, effort took experimental and additive forms.

In the morning we teach the curriculum normally, listening, speaking, reading, writing, communicatively, with role plays and presentations, and we even use AI chat tools for speaking. In the afternoon, students who chose English for the exam train for graduation. (Nancy, group interview)

Grace rebuilt her vocabulary teaching around the exam genres, saying "I teach vocabulary in exactly that context" (individual interview). Looking back, she described a path from initial surprise to growing confidence. Nancy credited her school leaders, who invited a university specialist to train teachers in using AI. Urban effort also reached the thinking beneath the items. Nancy saw reasoning as the main obstacle, since "even a student with the words cannot do the long gap-fill or the ordering task without logical thinking". She therefore built debates and recorded opinion clips to train it.

Rural teachers worked just as hard under tighter limits. Everlyn's department improvised because "the provincial subject council has held no training on how to teach" (group interview). For Alice, a 40-item paper in which every point counts pushed her teaching backwards.

Honestly I do not know what to do. I feel I am going back to the most ancient teaching method, Grammar Translation. I no longer dare to teach students skills like scanning or skimming. I make them translate everything. (Alice, group interview)

We do not read this retreat as unwillingness. Alice described herself as "a teacher who supports every change the Ministry makes". However, education authorities held workshops in the major cities, and "the school cannot fund attendance, I would pay myself, and that kills my motivation to go" (individual interview). Interestingly, one rural teacher was moving toward the construct on her own.

Students translated all four options, understood all of them, and still chose wrong, because the options mean nearly the same thing and differ only in what the writer wants to emphasise. So I am gradually shifting to a new method, teaching them, as in IELTS reading, to grasp the flow of the text and the writer's intention. (Everlyn, group interview)

Everlyn reached this point by trial and error. In our view, her case indicates that what rural teachers lacked was support, not insight.

4.3 Theme 3: Students who opted out of the English examination

The same reform, meeting unequal starting points, appeared to widen a gap that teachers already saw. In the rural case, the clearest sign was students choosing not to sit English once it became optional. Rose gave an estimate for her own school.

Out of about three hundred students in the cohort, only about forty will take English for graduation, and those forty are not all excellent, mostly around five or six points… It is very good for strong students. But if we look at the whole country, on a wide scale, I am certain it is not right at all. (Rose, group interview)

Alice similarly saw "no reaction" from her students, "only that they did not choose English for the exam" (group interview). Rose linked these choices both to the optional status of English and to the difficulty of the new paper. Urban teachers saw the widening from the advantaged side. Lily watched her class split on the day the sample paper arrived.

The students who study IELTS said, teacher, this suits me so well, I know it already. The ones who had not studied it said, why this wall of text, how terrifying. (Lily, group interview)

Jessie added that some schools give two English periods a week, while one big city's intensive programme gives six to eight (group interview). Grace observed that city parents "invest from when children are very small, systematically and financially" (individual interview).

Taken together, we find that the answer to RQ1 divides along geography and the answer to RQ2 along depth. Both cases complied with the format (Theme 1). The urban case turned that compliance into experimentation, whereas the rural case absorbed it as retreat (Theme 2). In turn, this difference shaped who continued to study English for the exam (Theme 3).

5 Discussion

5.1 Where the reform stood in Bailey's model (RQ1)

When we read the findings through Bailey's model [24], we see a reform that had reached some parts of the washback system much faster than others (Fig. 1). The test itself changed decisively, and the pathway from teachers to teaching was the most active one in our data, although it carried mainly a change of format. By contrast, the pathway from materials writers and curriculum designers to new materials lagged behind. Practice tests were plentiful, but teachers described no teaching materials built to develop the construct. With no results available in spring 2025, the feedback loop had not yet begun. We therefore place the reform at an early, top-down stage, in which the test and its training led while curriculum, materials and pedagogy followed slowly. Because the policy message reached classrooms mainly as sample papers and a matrix, we find it understandable that teachers treated the template as the message.

5.2 Technical washback and the depth of response (RQ2)

The responses we observed show change that was real and rapid but shallow. We propose the term technical washback for this pattern and define it as follows.

Technical washback is visible and effortful compliance with the surface of a reformed test, without engagement with the competence the test sets out to measure, during the implementation of the reform.

We see three features in the concept. First, it is effortful, which separates it from inertia. Second, its object is the item template, meaning the item types, genres and answer strategies that a sample paper makes visible. Third, it leaves the construct largely untaught. In Green's terms [25], the 2025 paper appears to have a favourable direction, since it overlaps with the curriculum's communicative goals far more than the old paper did. Intensity was also high in both settings. However, the depth of response stayed at the level of format. We therefore read the case as support for the argument we set out in Section 2.1, namely that overlap between test and construct may be necessary for positive washback but is unlikely to be sufficient.

Our data also point to three conditions that seem likely to produce technical washback. The first is a template signal, as the reform reached teachers mainly through sample papers and a matrix. The second is a short timeline, since teachers had months instead of years to prepare a cohort. The third is item-focused training, which signalled that matching the format was what preparation meant. Under these

conditions, we interpret technical washback less as a misunderstanding than as a sensible reading of the signals teachers received.

We recognise that the concept sits close to several existing ideas, and we see its value in what it adds to them. Teaching to the test describes instruction narrowed to test content, whereas technical washback concerns format and can occur even when a test samples its construct well. Coburn's [30] notion of shallow reform applies to any reform, while technical washback locates shallowness in teachers' engagement with a test construct. Vu and Nguyen [19] described superficial compliance with a curriculum, whereas an assessment arrives with its own concrete template. The pattern Wall [28] observed of new formats with old methods is perhaps the closest. Technical washback adds a definition and a place in a framework that separates it from both inertia and deeper change.

By crossing intensity with depth, we arrive at a working framework of four positions, which we call the Teacher Washback Response Framework (Fig. 2). We judged intensity from what teachers reported doing, such as rebuilding materials, converting school tests or adding classes. We judged depth from how they talked about the tasks, whether as formats and tips or as cohesion, coherence and genre. Technical washback combines high intensity with format-level engagement, and we found it in both cases. Substantive washback combines high intensity with engagement in the construct, and Grace's genre-based vocabulary teaching and Everlyn's turn toward text flow came close to it. Latent washback describes a teacher who recognises the construct without yet reorganising practice around it, as Jessie did. We did not find inert washback, where little changes, among the teachers themselves. However, Lily and Grace reported that their local education authorities still set mid-term and final tests in the old format, which points to inertia at the level of the system. We see the framework as a complement to Green's variability [25]. Variability explains why teachers differ, while the framework describes the forms those differences take along the dimension of depth.

The framework also sheds light on agency. In our reading, the urban dual track and the rural return to grammar translation were both judgements about what was possible in the moment [36]. What differed was the setting, which widened the choices available to urban teachers and narrowed those available to rural ones [35]. Importantly, Everlyn's discovery and Jessie's insight lead us to view technical washback as a state of the system, not a ceiling on teachers.

5.3 Implications for policy and teacher development

We draw three implications for the 2025 to 2035 agenda. First, we recommend that professional development under Decision 2371/QĐ-TTg [10] equip teachers with pedagogical techniques for teaching the construct, for example ways of making cohesion, coherence and genre visible to learners, and not only skills in writing items. Second, support needs to reach teachers who cannot travel to central training, because uniform cascade training appeared to reproduce the urban advantage. Third, school and provincial tests should carry the same construct as the national paper, since old-format local tests can hold teachers in place even when they are ready to move. In addition, we believe the optional status

of English deserves monitoring as an equity issue, because students who leave English for the exam may be those the repositioning agenda most needs to reach.

5.4 Directions for further research

We see this exploratory study as opening several directions. First, the 2026 paper kept the same six task types [43], which indicates that the format has stabilised. A second round of interviews, with the same teachers and new ones, could therefore trace how anticipated washback became enacted washback once results came out. Second, classroom observation could test how closely teachers' reported depth matches their practice. Third, survey research with larger samples, including teachers in mountainous and ethnic minority regions, could examine how widespread technical washback is. We offer three propositions from this study for such work to test. Washback is more likely to be technical when a new construct reaches teachers through sample papers on a short timeline. Item-focused training is likely to keep teachers in the technical position. Finally, the same initial compliance is likely to become experimentation where schools provide time and resources, and retreat where they do not.

6 Conclusion

The 2025 redesign of the NHGE English paper brought the test into line with the communicative goals of Vietnam's curriculum. For the teachers we interviewed, the reform changed the surface of every classroom. However, the teaching beneath appeared to change in proportion to local resources more than to policy intent. We have proposed technical washback as a way of naming this effortful but shallow change, together with a working framework that adds depth to the dimensions washback research already uses. Our case suggests that when the test changes first, teaching follows its format before its construct. In our view, closing that gap will depend on sustained, construct-focused support for teachers, which the 2025 to 2035 agenda still has time to build.

Declarations

Supplementary Information

The online version contains supplementary material (ESM), comprising the interview guide (Part A) and a comparison of task types and item counts across the 2024 paper, the October 2024 reference exam and the 2025 paper (Part B, Table S1).

Ethics approval and consent to participate

The study protocol was reviewed and approved in February 2025 by the Office of Science and Technology Management, Thanh Dong University, Hai Phong, Viet Nam, which serves as the university's institutional review body for research involving human participants. All procedures were performed in accordance with the institutional guidelines of Thanh Dong University. Informed consent was obtained

from all participants. All data were anonymised, stored securely and processed in accordance with Decree No. 13/2023/NĐ-CP on personal data protection by the Vietnamese Government.

Consent for publication

All participants consented to the publication of anonymised quotations.

Competing interests

The authors declare no competing interests.

Funding

The authors received no funding for this study.

Author Contribution

KNX designed the study, moderated the focus groups and one follow-up interview, coded the data and drafted the manuscript. TDNN conducted one follow-up interview, acted as a critical friend in the analysis and revised the manuscript. TVN acted as a critical friend in the analysis, reviewed the exam paper comparison and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgement

We thank the seven teachers who gave their time during a demanding examination season.

Data Availability

To protect participant confidentiality, we do not make the interview data publicly available. The corresponding author can provide anonymised extracts on reasonable request.

References

1. Alderson JC, Wall D. Does washback exist? Appl Linguist. 1993;14(2):115–. https://doi.org/10.1093/applin/14.2.115. 29.
2. Cheng L, Sun Y, Ma J. Review of washback research literature within Kane's argument-based validation framework. Lang Teach. 2015;48(4):436–70. https://doi.org/10.1017/S0261444815000233.
3. Parviz M. Reflecting on the consequences of the Iranian university entrance examination: a systematic-narrative hybrid literature review. Discov Educ. 2023;2:22.

https://doi.org/10.1007/s44217-023-00046-x.

4. Parviz M, Azizi M. The MSRT: a critical review of English proficiency in Iran. Discov Educ. 2025;4:226. https://doi.org/10.1007/s44217-025-00662-9.
5. Hankebo DD, Belayneh HB. Analyzing the washback effect of exit exam on the teaching and learning processes in law school at Wolaita Sodo University, Ethiopia. Discov Educ. 2026;5:226. https://doi.org/10.1007/s44217-026-01260-z.
6. Ministry of Education and Training. Đề thi tham khảo Kỳ thi tốt nghiệp THPT từ năm 2025 [Reference exam for the High School Graduation Examination from 2025]. 18 October 2024. https://xaydungchinhsach.chinhphu.vn/de-tham-khao-thi-tot-nghiep-thpt-mon-tieng-anh-tu-nam-2025-119241018160057549.htm
7. Ministry of Education and Training. Đề thi chính thức Kỳ thi tốt nghiệp THPT năm 2025, môn Tiếng Anh [Official 2025 NHGE English paper]. Reproduced by VietNamNet. 2025. https://vietnamnet.vn/de-thi-mon-tieng-anh-thi-tot-nghiep-thpt-nam-2025-2414634.html
8. Ministry of Education and Training. Quyết định số 4068/QĐ-BGDĐT phê duyệt Phương án tổ chức Kỳ thi và xét công nhận tốt nghiệp THPT từ năm 2025 [Decision No. 4068/QĐ-BGDĐT approving the examination plan from 2025]. 2023. https://luatvietnam.vn/giao-duc/quyet-dinh-4068-qd-bgddt-2023-cau-truc-de-thi-ky-thi-tot-nghiep-thpt-tu-2025-302643-d1.html
9. Communist Party of Vietnam. Kết luận số 91-KL/TW ngày 12/8/2024 của Bộ Chính trị về tiếp tục thực hiện Nghị quyết số 29-NQ/TW [Conclusion No. 91-KL/TW of the Politburo on continuing the fundamental renovation of education and training]. 2024. https://xaydungchinhsach.chinhphu.vn/toan-van-ket-luan-cua-bo-chinh-tri-ve-thuc-hien-doi-moi-can-ban-toan-dien-giao-duc-va-dao-tao-119240815133542026.htm
10. Vietnamese Government. Quyết định số 2371/QĐ-TTg phê duyệt Đề án. Đưa tiếng Anh thành ngôn ngữ thứ hai trong trường học giai đoạn 2025–2035, tầm nhìn đến năm 2045 [Decision No. 2371/QĐ-TTg approving the scheme on English as a second language in schools]. 2025. https://thuvienphapluat.vn/van-ban/Giao-duc/Quyet-dinh-2371-QD-TTg-2025-de-an-Dua-tieng-Anh-thanh-ngon-ngu-thu-hai-trong-truong-hoc-679020.aspx
11. Nguyen A. Mind the (theme) gap between assessment and curriculum: a mixed-method study on the National High School Examination of English in Vietnam from 2018 to 2020. Doctoral thesis. Milton Keynes: The Open University; 2020. https://oro.open.ac.uk/81095/
12. Ministry of Education and Training. Chương trình Giáo dục phổ thông 2018 [General Education Curriculum 2018]. Circular No. 32/2018/TT-BGDĐT. 2018. https://moet.gov.vn/tintuc/Pages/CT-GDPT-Moi.aspx?ItemID=8421
13. Ngo XM. English assessment in Vietnam: status quo, major tensions, and underlying ideological conflicts. Asian Engl. 2024;26(1):280–92. https://doi.org/10.1080/13488678.2022.2132128.
14. Watanabe Y. Teacher factors mediating washback. In: Cheng L, Watanabe Y, Curtis A, editors. Washback in language testing: research contexts and methods. Mahwah: Lawrence Erlbaum Associates; 2004. pp. 129–46.

15. Khan A, Hassan N, Cheng L. Investigating the contextual factors mediating washback effects of a learning-oriented English language assessment in Malaysia. Lang Test Asia. 2025;15(1):20. https://doi.org/10.1186/s40468-025-00359-8.
16. Liang Y, Siraj S, Lo YY. Exploring washback processes in language teaching through teachers' beliefs in the Chinese National Matriculation English Test (NMET). Humanit Soc Sci Commun. 2025;12:1127. https://doi.org/10.1057/s41599-025-05345-2.
17. Dao TTD, Ngo PLH, Chi DN. Teacher agency and curriculum changes: a case study of two English teachers in Vietnam and their implementation of the new textbooks policy at the classroom level. TESOL J. 2025;16(3):e70055. https://doi.org/10.1002/tesj.70055.
18. Nguyen HTM, Bui T. Teachers' agency and the enactment of educational reform in Vietnam. Curr Issues Lang Plan. 2016;17(1):88–105. https://doi.org/10.1080/14664208.2016.1125664.
19. Vu T, Nguyen LTH. Navigating English curriculum reform in Vietnam: a socio-cultural perspective on teacher perezhivanie and agency. Curric Perspect. 2026;46(1):35–48. https://doi.org/10.1007/s41297-025-00315-9.
20. Nguyen CLT, Dao QTN. Assessing the washback effects of the 2025 National High School Graduation Exam: a cross-sectional study in Vietnam. Vietnam J Educ. 2025;9(3):424–36. https://doi.org/10.52296/vje.2025.565.
21. Le VC. English language teaching in Vietnam: aspirations, realities, and challenges. In: Le VC, Nguyen HTM, Nguyen TTM, Barnard R, editors. Building teacher capacity in English language teaching in Vietnam: research, policy and practice. London: Routledge; 2019. pp. 7–22. https://doi.org/10.4324/9780429457371-2.
22. Messick S. Validity and washback in language testing. Lang Test. 1996;13(3):241–56. https://doi.org/10.1177/026553229601300302.
23. Alderson JC, Hamp-Lyons L. TOEFL preparation courses: a study of washback. Lang Test. 1996;13(3):280–97. https://doi.org/10.1177/026553229601300304.
24. Bailey KM. Working for washback: a review of the washback concept in language testing. Lang Test. 1996;13(3):257–79. https://doi.org/10.1177/026553229601300303.
25. Green A. IELTS washback in context: preparation for academic writing in higher education. Studies in Language Testing 25. Cambridge: Cambridge University Press; 2007.
26. Shih CM. How tests change teaching: a model for reference. Engl Teach Pract Crit. 2009;8(2):188–206.
27. Spratt M. Washback and the classroom: the implications for teaching and learning of studies of washback from exams. Lang Teach Res. 2005;9(1):5–29.
28. Wall D. The impact of high-stakes examinations on classroom teaching: a case study using insights from testing and innovation theory. Cambridge: Cambridge University Press; 2005.
29. Qi L. Stakeholders' conflicting aims undermine the washback function of a high-stakes test. Lang Test. 2005;22(2):142–73. https://doi.org/10.1191/0265532205lt300oa.

30. Coburn CE. Rethinking scale: moving beyond numbers to deep and lasting change. Educ Res. 2003;32(6):3–12. https://doi.org/10.3102/0013189X032006003.
31. Canale M, Swain M. Theoretical bases of communicative approaches to second language teaching and testing. Appl Linguist. 1980;1(1):1–47. https://doi.org/10.1093/applin/1.1.1.
32. Celce-Murcia M, Dörnyei Z, Thurrell S. Communicative competence: a pedagogically motivated model with content specifications. Issues Appl Linguist. 1995;6(2):5–35. https://doi.org/10.5070/L462005216.
33. National Assembly of Vietnam. Nghị quyết số 88/2014/QH13 về đổi mới chương trình, sách giáo khoa giáo dục phổ thông [Resolution No. 88/2014/QH13 on renewing general education curricula and textbooks]. 2014. https://thuvienphapluat.vn/van-ban/Giao-duc/Nghi-quyet-88-2014-QH13-doi-moi-chuong-trinh-sach-giao-khoa-giao-duc-pho-thong-260798.aspx
34. Ministry of Education and Training. Quyết định số 3588/QĐ-BGDĐT về việc lựa chọn bộ sách giáo khoa sử dụng thống nhất toàn quốc [Decision No. 3588/QĐ-BGDĐT on the textbook series for unified nationwide use]. Reported by Báo Chính phủ. 26 December 2025. https://baochinhphu.vn/ket-noi-tri-thuc-voi-cuoc-song-la-sach-giao-khoa-thong-nhat-toan-quoc-tu-nam-hoc-20262027-10225122622052997.htm
35. Priestley M, Biesta G, Robinson S. Teacher agency: an ecological approach. London: Bloomsbury; 2015.
36. Emirbayer M, Mische A. What is agency? Am J Sociol. 1998;103(4):962–1023. https://doi.org/10.1086/231294.
37. Tran H, Li M. Influential sociocultural factors on teacher agency in times of educational change: reflection from a Southeast Asian context. Int J Appl Linguist. 2025;35(1):308–24. https://doi.org/10.1111/ijal.12619.
38. Stake RE. The art of case study research. Thousand Oaks: SAGE; 1995.
39. Braun V, Clarke V. Thematic analysis: a practical guide. London: SAGE; 2021.
40. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res. 2016;26(13):1753–60. https://doi.org/10.1177/1049732315617444.
41. Lambert SD, Loiselle CG. Combining individual interviews and focus groups to enhance data richness. J Adv Nurs. 2008;62(2):228–. https://doi.org/10.1111/j.1365-2648.2007.04559.x. 37.
42. Smith B, McGannon KR. Developing rigor in qualitative research: problems and opportunities within sport and exercise psychology. Int Rev Sport Exerc Psychol. 2018;11(1):101–21. https://doi.org/10.1080/1750984X.2017.1317357.
43. Ministry of Education and Training. Đề thi chính thức Kỳ thi tốt nghiệp THPT năm 2026, môn Tiếng Anh, mã đề 1116 [Official 2026 NHGE English paper, code 1116]. 2026. https://xaydungchinhsach.chinhphu.vn/thi-tot-nghiep-thpt-2026-de-thi-tieng-anh-va-goi-y-dap-an-119260612132828853.htm

Figures

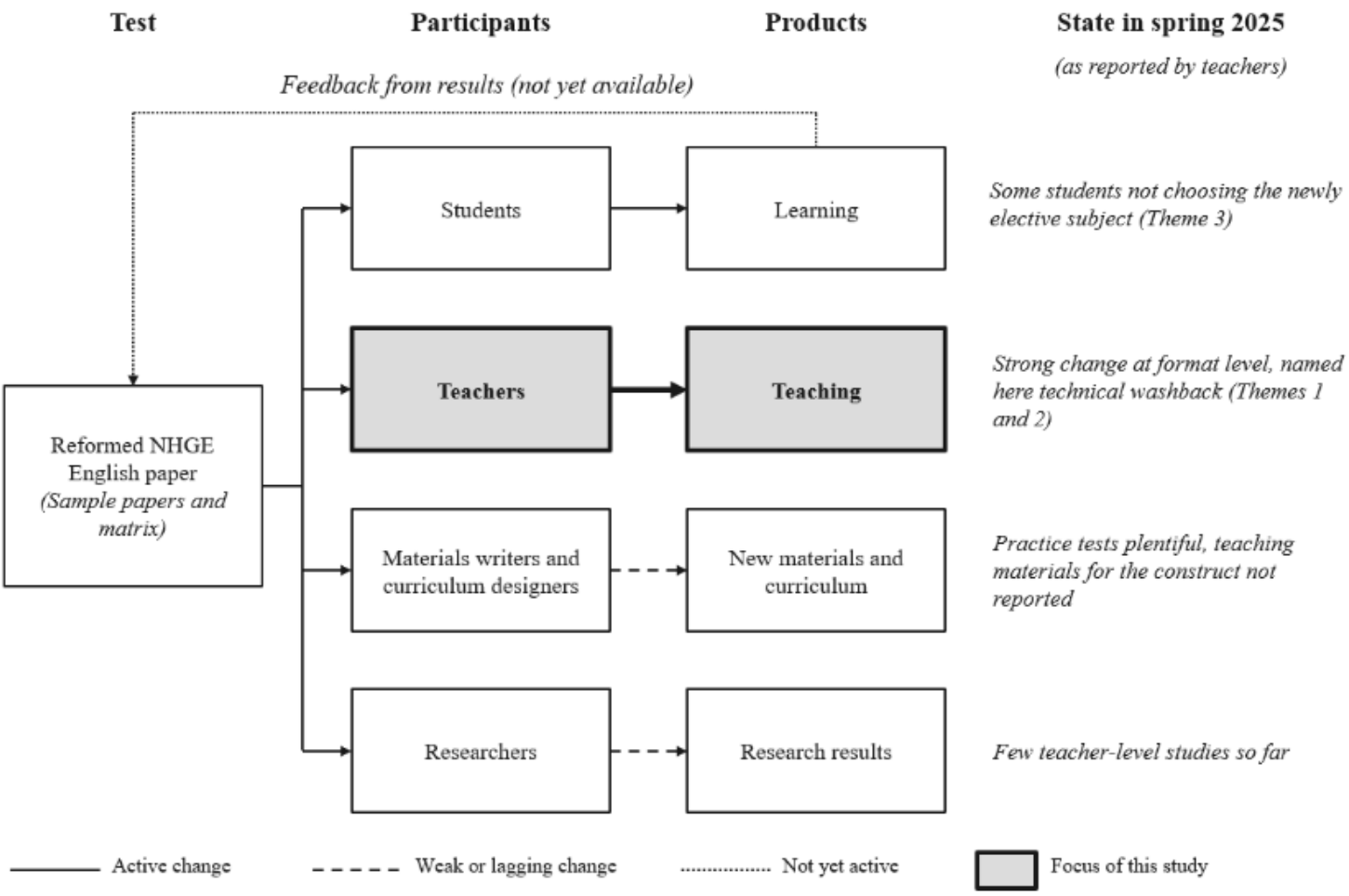


Figure 1

Bailey's basic model of washback adapted to the 2025 NHGE implementation phase. Line styles show the state of each pathway in spring 2025, as teachers reported it and as the exam documents show it

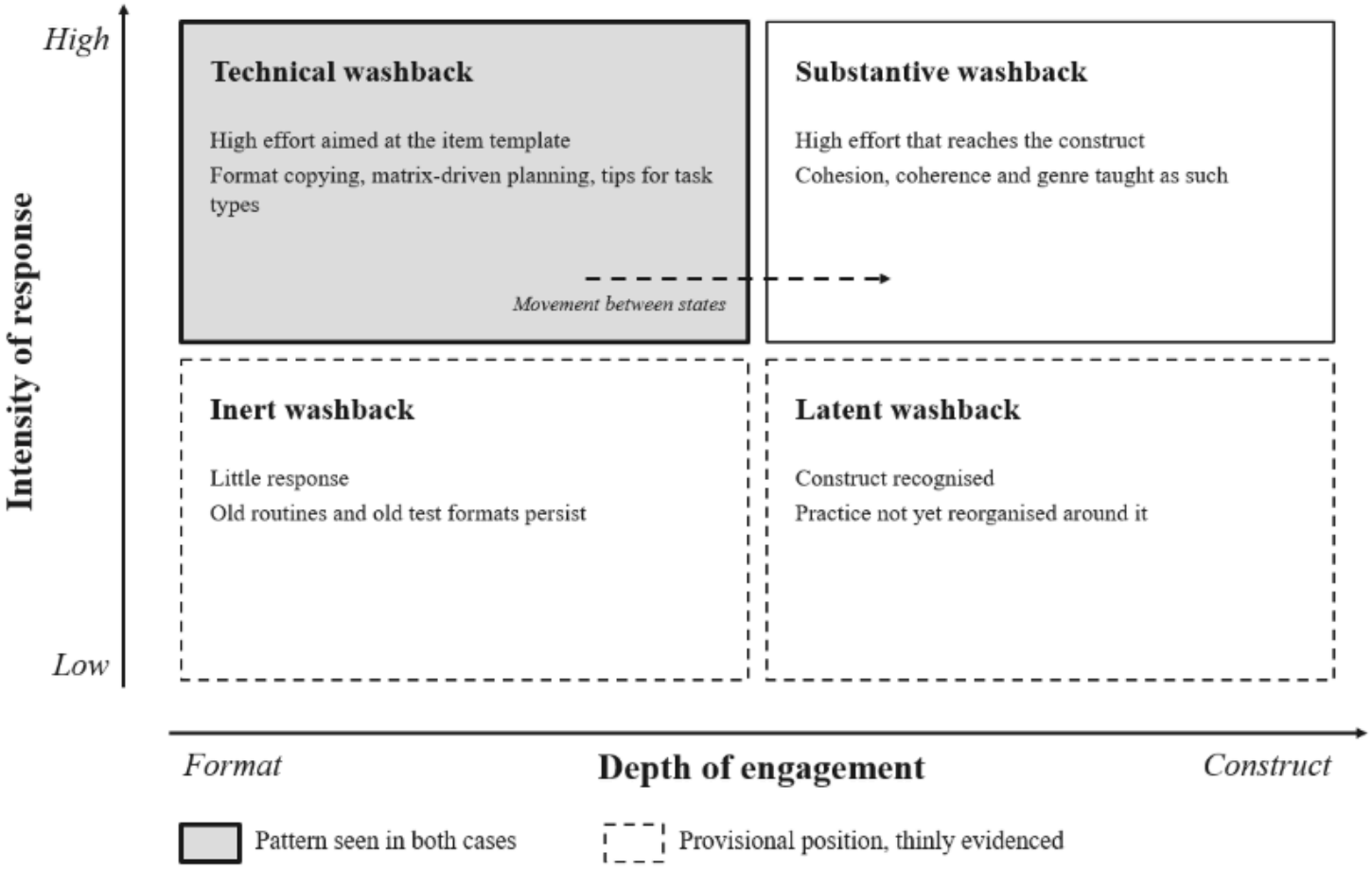


Figure 2

Teacher Washback Response Framework. We developed the framework from our analysis of the interview data. It places a teacher's response according to how much effort it involves and how far it reaches into the test construct. The four positions describe practice at one moment, not types of teachers, and a teacher can move from one position to another. Shading marks the position we found in both cases, and dashed borders mark the two positions that our data illustrate least fully

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESMDiscoverEducation.docx