Language Testing in Asia

RESEARCH Open Access

The washback of the National Matriculation English Test on senior high school English learning outcomes: do test takers from different provinces think alike?

Hao Zhang[1*]

*Correspondence:
zhanghao_SFLL@bnu.edu.cn

[1] School of Foreign Languages and Literature, Beijing Normal University, Haidian District, No.19 Xinjiekouwai Street, Beijing 100875, China

Abstract

The Chinese National Matriculation English Test (NMET) has a test purpose of producing beneficial washback to promote senior high school English teaching and learning. This article presents a large-scale nationwide survey research on student perceptions of the NMET's before-test washback on their English learning outcomes in senior high school and the influence of the factor of home province on their views. The research participants were 20,062 first-year undergraduates from 103 universities in the Chinese mainland, and findings showed that the NMET facilitated their senior high school English learning outcomes in the surveyed skills both tested and not tested in the NMET. However, there were noticeable differences between students' expectations and realities. It was also found that home province was a significant predictor of student perceptions, though the strengths of the relationships varied. Furthermore, in order for more beneficial NMET washback, the students suggested making the NMET speaking and listening subtests mandatory across all provincial regions and administering a nationally unified NMET form. This study calls on future researchers to hold a close lens to the complexity and dynamism of the NMET washback over time and across situations.

Keywords: Washback, The National Matriculation English Test (NMET), Student perspective, Senior high school English learning, Home province

Introduction

In the Chinese mainland, the *gaokao* (University Entrance Examination to Higher Education), including the National Matriculation English Test (NMET), is nicknamed the *footslog bridge*—the only route to success in study, work and life (Cheng & Curtis, 2010). It is a high-stake examination since its results create "winners and losers, successes and failures, rejections and acceptances" (Shohamy, 2000, p. 2). The combination of *gaokao*'s far-reaching consequences on test takers' future and its substantial impact on education and society has created a "highly-valued, highly-selective, rather narrowly-defined examination-driven context" (Cheng & Curtis, 2010, p. 267) which has shaped Chinese people's lives for generations.

Toward the end of their senior high school education, students sign up and sit for the *gaokao* in their home provinces where their household registration, or *hukou*, is recorded. All *gaokao* candidates must take a foreign language test and the NMET is chosen by over 99% of *gaokao* test takers each year[1] (Liu, 2010). The annually released NMET test manual (e.g., MoE, 2022) draws guidance from the *English Curriculum Standards for General Senior High Schools (2017 Edition)* (MoE, 2018), the *Gaokao Assessment Framework* (NEEA, 2019), and *China's Standards of English Language Ability* ([CSE], MoE & State Language Affairs Commission, 2018)—(unified, scaled, and use-oriented English proficiency criterion for all levels of English education in the Chinese mainland. Test takers' NMET scores, together with those of several other subjects, are tallied in their total *gaokao* scores used to make admissions decisions for higher educational programs.

The NMET aims to (1) facilitate moral education, (2) select qualified senior high school graduates for tertiary education, and (3) produce positive washback on secondary English education (National Education Examinations Authority of China [NEEA], 2019), specifically, to bring about changes "from formal linguistic knowledge to practice and use of the language" (Li, 1990, p. 402). According to the NMET test manual (MoE, 2022), the test examines test takers' English abilities in reading, language usage, writing, listening, and speaking. The reading, language usage, and writing components are delivered in a pen-and-paper manner. In 2022, 27 of the 31 provincial areas used one of the four NMET pen-and-paper forms developed by the NEEA—the MoE-appointed institution exclusively supervising educational examinations and exerting administrative authority, while the remaining four (i.e., Beijing, Shanghai, Tianjin, and Zhejiang) designed their own tests. It is also specified in the test manual that all NMET forms, developed by the NEEA or not, are constructed in compliance with the *English Curriculum Standards for General Senior High Schools (2017 Edition)* (MoE, 2018) and the *Gaokao Assessment Framework* (NEEA, 2019). Therefore, despite the parallel NMET forms for different provincial areas, the test policies, specifications, scope, content, format, and administration of the NMET remain basically consistent across the Chinese mainland (see Table 1 for an example of the NMET pen-and-paper test format).

Until recent years the listening component was also administered pen-and-paper and the speaking subtest face-to-face, while in 2022, Beijing, Guangdong, and Shanghai delivered computer-based listening-speaking integrated subtests. English reading, language usage, and writing all count toward college[2] admissions across the Chinese mainland, but listening and speaking are not compulsory in all provincial regions. In 2022, seven provincial areas—Gansu, Heilongjiang, Henan, Inner Mongolia, Liaoning, Shanxi, and Xinjiang—did not add test takers' listening scores in their total NMET results or even did not deliver a listening subtest, while only Beijing, Guangdong, and Shanghai required all their *gaokao* candidates to sit for the speaking component.

[1] The other foreign language tests *gaokao* candidates can choose are Russian, Japanese, German, French, and Spanish.

[2] "College" and "university" are used interchangeably in this article to refer to post-secondary education or a higher education institution.

Table 1 Test format of the NEEA-developed 2022 NMET Versions I and II (pen and paper)

Section	Part	Input	Language of rubric	Item types	Number of items	Raw score	Percentage
I. Listening	A	Five short dialogues	Chinese	3-option multiple choice (MC)	5	7.5	5%
	B	Five dialogues or monologues	Chinese	3-option MC	15	22.5	15%
II. Reading Comprehension	A	Four texts	Chinese	4-option MC	15	30	20%
	B	One text	Chinese	7-option gap-filling	5	10	≈6.7%
III. English Language Usage	A	One text	Chinese	4-option MC cloze	20	30	20%
	B	One text	Chinese	Word transformation	10	15	10%
IV. Writing	A	One text	Chinese	Error correction	10	10	≈6.7%
	B	Prompt	Chinese	Guided writing	1	25	≈16.6%
Total					80 + 1	150	100%

The NMET has been criticized for its alleged failure in producing positive beneficial washback on secondary English education (e.g., Wang & Zhan, 2011; Zhu & Yang, 2004). The test has been blamed for demoting secondary English education to a de facto examination-oriented machine for raising students' NMET scores without paying much attention to improving students' English language ability in use (Zhu & Yang, 2004), the repercussion of which is *teaching to the test* (Madaus, 1988). However, these criticisms are more often not empirically demonstrated.

Even though the NEEA has not published extensive official documents on NMET quality controls, academic literature has reported that measures have been taken to improve the validity (e.g., Chen et al. 2019), reliability (e.g., Li, 1990), discrimination power (e.g., Cheng & Qi, 2006), and fairness (e.g., Luo & Xiao, 2018; Zhang, 2019) of the NMET. Raising these test qualities has made the NMET an effective means to screen and select senior high school graduates with sufficient English proficiency for tertiary education. On the other hand, another NMET test purpose—re create positive washback on senior high school English teaching and learning—pos yet to receive the attention it merits (Qi, 2011).

Only a few small-scale empirical studies have explored the washback of NMET on secondary English education (e.g., Li, 1990; Qi, 2004), not to mention nationwide investigations. Additionally, most of these studies have predominantly focused on NMET washback on teaching practices, with limited attention given to that on learning processes (e.g., Qi, 2004), while scarce research has examined test takers' perspectives upon the washback effect of NMET on their senior high school English learning outcomes. Consequently, this aspect remains underexplored in the current literature. Moreover, previous research has paid little attention to inter-province differences in NMET washback on secondary English education. One of the goals of the NMET is to create positive washback on both English teaching and learning in senior high schools

(NEEA, 2019). As such, the test can impact English curriculum design, pedagogical approaches, and learning practices in secondary schools (Gui et al. 1988). In addition, since not all provincial areas use the same NMET form and wide social, economic, and educational divergence exists across the Chinese mainland, the washback effect of NMET on secondary English education may vary between provincial regions. As a result, we conducted a large-scale nationwide investigation to survey post-secondary students about the NMET washback on their English learning outcomes in senior high school.

Literature review

In field of language testing and assessment, the niche of washback studies was officially established by Alderson (1986). Alderson and Wall (1993) christened *washback* as the extent to which the power of a test will lead teachers and learners to do what "they would not necessarily otherwise do because of the test" (p. 117). It refers to the effects of a test on teaching and learning in educational systems (Bachman & Palmer, 2010). Alderson and Wall's (1993) and Alderson and Hamp-Lyons' (1996) seminal publications offered altogether 16 washback hypotheses. These hypotheses can be classified into (1) washback on teaching; (2) washback on learning; and (3) amounts of washback. The first category includes washback effect on teaching content, methods, rate, sequence, depth, and attitude. Hypotheses about learning involve washback on learning content, strategies, rate, sequence, depth, and attitude, while hypotheses related to amounts of washback concern the strength of the washback and indicate that a test's stakes may also affect its washback effect. In the meantime, not all learners and other stakeholders as well experience the same type and/or amounts of washback. These hypotheses laid a solid research groundwork and set the course for future washback studies. The present study is no exception in this regard. The washback hypotheses on learning guided part of our research design.

Watanabe (1997, 2004a) held that the complicated nature of washback could be conceptualized from five dimensions—specificity, intensity, length, intentionality, and value :

- *Specificity* refers to whether the washback is generated by all tests (general washback) or by one particular aspect of a test or one specific type of tests (specific washback).
- *Intensity* is the degree of test washback on teaching and learning (strong or weak washback).
- *Length* concerns whether the washback lasts for a short or long period of time. If the washback of a test is found existing before test takers sit the test, that is before-test short-term washback; yet if the test continues to influence test takers after they have taken the test, then this test has after-test long-term washback.
- *Intentionality* relates to whether the washback is intended or unintended by test constructors and testing agencies (intended or unintended washback).
- *Value* refers to whether a test's washback facilitates or hinders teaching and learning (positive or negative washback).

The present research pertains to one specific type of tests—national higher education admissions English test, in the context of the Chinese mainland. The scope of the research focuses on the before-test short-term washback in that one of the research aims is to explore whether the high-stake NMET has achieved one of its intended goals to produce strong positive washback on students' English learning in senior high schools.

From another perspective, Hughes (1993) discussed the operating washback trichotomy of participants, processes, and products in teaching and learning. Integrating this trichotomy with Alderson and Wall's (1993) washback hypotheses, Bailey (1996) proposed her model of washback that suggests a cyclical and systematic way of understanding washback—tests wielding direct influences on the *participants* involved in various *processes* and these processes yield different *products* to different participant groups, and finally all the products loop back onto the tests. This model improved the washback hypotheses by including in the washback mechanism a wider range of stakeholder groups than only teachers and learners, and by considering bi- or multi-directional rather than linear relationships between a test and its washback on teaching and learning. However, Bailey (1996) did not specify the *processes* in her washback model, and an effort in this line is Green's (2003, 2006, 2007) predictive model of test washback.

Green's (2003, 2006, 2007) model focused on what might happen in the *processes* that would produce varied washback effects. This model comprised three modules (i.e., washback direction, variability, and intensity) and underlined a deep understanding of the test itself and how the test construct influences its washback. It also took into account the relationship between test design and the local testing context. Another feature of Green's (2003, 2006, 2007) model is that it expanded the washback mechanism to incorporate a domain of test design and emphasized the relationship between stakeholders' perceptions of a test (e.g., its importance, stakes, and difficulty) and the intensity of its washback effects.

These washback hypotheses and models have inspired a substantial amount of empirical studies on the washback effects of specific language tests, especially three groups of English tests: (1) international proficiency tests; (2) national/regional proficiency test; and (3) national/regional matriculation tests. Most previous washback research on international English proficiency tests focused on the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) (see Tables 2 and 3 for summaries). Most of these studies covered washback on teaching (e.g., teachers' teaching methods, materials, attitudes toward the test) and learning (e.g., students' learning processes, outcomes, attitudes toward the test, preparation strategies).

Another line of empirical washback studies is on national or regional English proficiency test, particularly the General English Proficiency Test (GEPT), the College English Test (CET), and the Test for English Majors (TEM) (see Table 4 for a summary). Shih (2006, 2007, 2009, 2010) found different degrees of GEPT washback on stakeholders and students did not invest much effort in test preparation. Similarly, Vongpumivitch's (2012) findings also showed mainly negative GEPT washback. Regarding the CET, Jin (2000) revealed that the speaking subtest of the CET produced positive washback on college English teaching in the Chinese mainland. Later, Gu's (2007, 2014) systematic research demonstrated that positive CET washback on college English teaching outweighed its negative effect. Unfortunately, these CET washback

Table 2 Research on IELTS washback

Researcher(s) (Year of publication)	Geographical context	Research method(s)	Participants	Major findings
Coleman et al. (2003)	Australia; UK; The Chinese mainland	Questionnaire; interview	Test takers; university staff	- IELTS was perceived to have high validity; - Students were satisfied with IELTS entry scores and their English abilities - University Teachers wished for an increased IELTS entry score and were unsatisfied with students' English abilities
Hayes and Read (2004)	New Zealand	Questionnaire; classroom observation; interview; testing	Teachers; students	- Teaching-to-the test was found in the focal IELTS preparation course; - Great differences between IELTS preparations courses and EAP courses; - No significant improvement of students' IELTS scores was found after they took the focal preparation course

Hawkey (2006)	Not specified	Questionnaire; classroom observation; interview; analysis of textbooks and materials	Teachers;
			test takers;
			IELTS administrators;
			admission officers
				Participants generally perceived that:
				- IELTS tests all four communicative macro-skills;
				- IELTS is a fair, authentic and high-stake test;
				- Reading and writing modules are more difficult

Green (2006)	UK	Classroom observation; interview	Teachers	- Common practices were found in both IELTS and EAP classes; - Great differences exist between IELTS and EAP classes - Teacher variables induced practices not predicted from the test design

studies did not address CET washback on students' learning. Ma (2019) also reported positive CET washback on students' test preparation. More attention was also given to washback on teaching than on learning in often-cited TEM washback studies. Xu (2012) and Zou and Xu (2017), based on the same dataset, found generally satisfactory positive TEM washback on English teaching, while Zou and Xu (2014) touched upon washback on learning by reporting that the TEM increased students' learning motivation.

Table 3 Research on TOEFL washback

Researcher(s) (Year of publication)	Geographical context	Research method(s)	Participants	Major findings
Alderson and Hamp-Lyons (1996)	USA	Classroom observation; interview	Teachers; students	- TOEFL affected both teaching content and methods - Teachers were affected by washback of different strength and type - TOEFL alone did not cause washback
Wall and Horák (2006)	Central and Eastern Europe	Classroom observation; interview	Teachers; students; directors of studies	- a hope for the new TOEFL test to induce a more communicative teaching approach - a hope for TOEFL prep classes to include academic tasks; - TOEFL prep classes were coursebook-based and teacher-dominated
Wall and Horák (2008)	Central and Eastern Europe	Interview	Teachers	- Teachers' attitudes toward the new TOEFL test were generally positive - Their understandings of the differences between the new and old TOEFL tests increased with time - They were not sure how to incorporate the new TOEFL into their teaching
Wall and Horák (2011)	Central, Eastern and Western Europe	Classroom observation; interview; analysis of coursebooks	Teachers	- A strong influence from TOEFL coursebooks on course design and teaching methods - Clear changes in teaching reading, listening, speaking, integrated writing and grammar after the introduction of the new TOEFL test, but no change in teaching independent writing; - The major washback of the new TOEFL was on the teaching content

The third type of English tests whose washback has been actively investigated is national or regional matriculation tests, such as those in Sri Lanka (e.g., Wall & Alderson, 1993), Israel (e.g., Ferman, 2004; Shohamy, 1993; Shohamy et al. 1996), Japan (e.g., Watanabe, 1996, 2004b), Hong Kong, China (e.g., Cheng, 2005; Cheng et al. 2011), and the Chinese mainland (e.g., Li, 1990; Qi, 2004; Zhang & Bournot-Trites, 2021) (see

Table 4 Research on washback of national/regional English proficiency tests

Target test	Researcher(s) (Year of publication)	Geographical context	Research method(s)	Participants	Major findings
The General English Proficiency Test (GEPT)	Shih (2006)	Taiwan	Classroom observation; interview; document review	Teachers; students; students' family members; university department chairs	- The GEPT had limited washback on teaching at both universities; - Teachers held different views toward the validity and reliability of the GEPT while students believed that the GEPT had gained public credibility; - Existing theories and models could not fully explain GEPT's washback on learning, teaching or departmental and school policies
	Vongpumivitch (2012)	Taiwan	Questionnaire	Test takers	- Only a slight majority of participants thought the GEPT had motivated them to learn English; - Most participants did not think GEPT was successful in promoting lifelong learning
The College English Test (CET)	Jin (2000)	The Chinese mainland	Questionnaire	Test takers; examiners	- The validity and reliability of the CET-SET were perceived satisfactory; - The CET-SET had positive washback on college English teaching and learning in China
	Gu (2007)	The Chinese mainland	Questionnaire; classroom observation; interview; tests; analyses of relevant materials and data	Teachers; students; administrators	- Most participants thought highly of CET; - The CET had both positive and negative washback on college English teaching but the strength varied; - Creative uses of textbooks and other teaching materials led to higher CET scores

Table 4 (continued)

Target test	Researcher(s) (Year of publication)	Geographical context	Research method(s)	Participants	Major findings
	Gu (2014)	The Chinese mainland	Questionnaire; classroom observation; interview; journals; video recording	Teachers; students	- The CET brought changes to teaching plans and content; - The positive washback of CET outweighed its negative effects
	Ma (2019)	The Chinese mainland	Questionnaire; an official test	Students	- The CET 4 had positive washback on students' test preparation; - Significant relationships were identified between students' characteristics and CET scores
The Test for English Majors (TEM)	Xu (2012); Zou and Xu (2017)	The Chinese mainland	Questionnaire	Foreign language experts; English discipline leaders	- Participants were generally positive toward TEM 8 and its washback; - The limited useful information in the test report and low transparency of the marking criteria caused negative washback
	Zou and Xu (2014)	The Chinese mainland	Questionnaire	Teachers; test takers; university program administrators	- Different stakeholders held different views toward the various aspects of TEM; - Stakeholders tended to use TEM results for non-instructional purposes; - Most participants were satisfied with the TEM and its washback on teaching and students' learning motivation

Table 5 Research on washback of national/regional matriculation English tests

Geographical context	Researcher(s) (Year of publication)	Target test	Research method(s)	Participants	Major findings
Sri Lanka	Wall and Alderson (1993)	The O-level English exam	Questionnaire; classroom observation; interview	Teachers; students	- Both positive and negative washback effects were found on teaching content but not on teaching methods
Israel	Shohamy (1993)	The ASL Test; The EFL Oral Test; The L1 Reading Comprehension Test	Questionnaire; classroom observation; interview; documents analysis	Teachers; students	- Strong washback from all three tests; - Both positive and negative washback on teaching content, methods, and materials; - Teaching-to-the-test abounded in instruction
	Shohamy et al. (1996)	The ASL; The EFL Test;	Questionnaire; interview; documents analysis	Teachers; students; EFL inspectors	- Increasing EFL washback but decreasing ASL washback; - Washback effects of the tests changed over time
	Ferman (2004)	The EFL Oral Matriculation Test	Questionnaire; interview; documents analysis	Teachers; students; EFL inspectors	- Different washback on different students; - The test exerted both positive and negative washback on teaching and learning;
Japan	Watanabe (1996)	The Japanese university entrance examination	Classroom observation; interview	Teachers	- Teacher factors such as educational background, personal beliefs and teaching experience had more influence on washback than the test itself
	Watanabe (2004b)	The Japanese university entrance examination	Classroom observation; interview	Teachers	- Teachers' concerns for students' abilities, their perceptions of the exam and their familiarity with various teaching methods played important parts in the process of engineering washback

Table 5 (continued)

Geographical context	Researcher(s) (Year of publication)	Target test	Research method(s)	Participants	Major findings
Hong Kong, China	Cheng (2005)	The HKCEE	Classroom observation; questionnaire; interview	Teachers; students; examiners; book publishers; school administrators	- Efficient washback on teaching content but not on teaching methods; - The exam had washback effects on both macro and micro levels of the educational system
	Cheng et al. (2011)	The SBA in the HKCEE	Questionnaire	Students; students' parents	- A correlation between students' attitudes of test preparation activities and their English ability; - Certain variables affected parents' perceptions
The Chinese mainland	Li (1990)	The MET	Questionnaire	Teachers; local English inspectors	- positive washback on in-class teaching content and materials; - After-class learning changed considerably
	Qi (2004)	The NMET	Classroom observation; questionnaire; interview	Teachers; students; English inspectors; NMET test constructors	- NMET could not bring its intended washback because of a conflict between its two functions; - The Senior III English course was dedicated to NMET preparation
	Zhang and Bournot-Trites (2021)	The NMET	Questionnaire; interview	Students	- The NMET had generally positive long-term washback on their college English learning outcomes; - More NMET preparation in terms of English speaking and listening skills would have better prepared them for undergraduate study; - Participant perceptions were associated with province of origin, senior high school status, and college status

Table 5 for a summary). The following only reviews the NMET washback research in the Chinese mainland that is immediately relevant to our study.

Li's (1990) study is to our knowledge the earliest published empirical NMET (then called MET) washback study. From the perspective of teachers and local English teaching-and-research inspectors, the MET generated intended washback effects, to different degrees, on English teaching and learning in secondary education. A shift was palpable from merely teaching language knowledge to combing knowledge teaching with skill training in reading, listening, speaking, and writing. Also noticeable was the change to after-class learning. Since both nationally unified textbooks and imported as well as self-compiled materials became available, students seemed to have developed stronger awareness of time and resources and greater passion for after-class English learning. Similar to most previous studies, Li's (1990) research did not involve student participants, and its main focus was on washback on English teaching rather than on learning.

Continuing the theme of NMET washback on secondary English education, Qi (2004) collected both quantitative and qualitative data from a wider range of stakeholders (i.e., NMET test constructors, local English inspectors, Senior III teachers and Senior III students). Specifically, this study surveyed NMET washback on teaching material selection, teaching focuses, teaching and learning activities, testing and assessment methods, learning motivations, learning aims, learning attitudes and learning strategies. Qi's (2004) findings showed that the NMET did not produce the intended washback, and this was for the most part due to the inherent conflict between the NMET's purposes to screen students for higher education and to bring about beneficial changes to secondary education. It was also found that Senior III English lessons were in fact coaching sessions devoted to doing mock tests, repeatedly reviewing formal linguistic knowledge, and mechanically drilling the language skills tested in the NMET. Additionally, non-test factors also played a role in Senior III English teaching—teachers' English proficiency, their own learning experience, potential misuses of NMET test score, etc. In light of these findings, Qi (2004) concluded that only by lowering the stakes of the NMET to an appropriate level can the test effectively creates intended washback.

Although Qi's (2004) research was comprehensive and detailed, it was conducted two decades ago. Thereafter, the NMET went through five rounds of reform, and its test format and coverage are not the same as they used to be back in the 2000s. Further, Qi's (2004) study only recruited participants from two provinces and therefore was not able to provide a panorama of NMET washback across the large territory of the Chinese mainland. Furthermore, this study explored NMET washback on students' learning motivations, aims, attitudes, and strategies— ai learning processes as defined by Bailey (1996), but it did not investigate washback on students' learning outcomes/products which are integral to Hughes' (1993) washback mechanism and subsequently Bailey's (1996) washback model.

As an effort to fill one of the gaps left by Qi (2004), Zhang and Bournot-Trites' (2021) study focused on NMET washback on students' learning outcomes. They used questionnaires and interviews to examine long-term NMET washback effects on tertiary students' English learning outcomes in higher education institutions. They found that despite students' generally positive evaluations of NMET washback on their college English learning outcomes, NMET preparation in senior high school was comparatively not adequate at preparing students with sufficient listening and speaking skills for higher education. This was partly attributable to the status quo that the NMET listening and speaking subtests were not mandatory in all provincial regions across Chinese mainland. The authors suggested that the NMET's long-term washback on tertiary English education is a continuation and extension of its short-term washback on secondary English learning, and it would help smooth transition for students from senior high school to undergraduate study if the NMET included a test purpose to also produce beneficial washback on college English learning, rather than only on senior high school English education as was the case then.

To our knowledge, Zhang and Bournot-Trites' (2021) research has been to date the only detailed account of long-term NMET washback on Chinese EFL students' college English learning outcomes, and it contributed to a fuller understanding of NMET washback mechanism in the Chinese context. Having said that, their study concentrated on the NMET's after-test washback on college English learning, while short-term before-test NMET washback on senior secondary students' English learning outcomes—the antecedent of its long-term washback and an important gap left by prior studies—has yet to be systematically explored. Moreover, Zhang and Bournot-Trites (2021) was a small-scale investigation, and though they sampled more widely across the Chinese mainland than Qi (2004), extrapolations to the state level were limited and results might be subject to interpretation.

Overall, three decades of substantial washback studies have proposed and tested a wide range of conceptual and theoretical washback frameworks that have revealed the multifaceted and complex nature of washback in different contexts. In respect of the NMET washback, the contributions of previous research have included short-term washback on test preparation and senior high school English teaching, and long-term washback on college English learning. There is a need for research on short-term washback on students' before-NMET English learning in senior high school. Besides, given the small-sampling nature of prior NMET washback studies, additional compelling evidence is necessary for a general picture and in-depth understanding of NMET washback across the vast territory of the Chinese mainland with considerable socioeconomic and educational variations between provincial regions. Given that the NMET serves the same research purposes across the Chinese mainland, irrespective of provincial boundaries, it is justified to conduct inter-province comparisons to examine to what extent the NMET has achieved one of its objectives across provinces to facilitate senior high school English learning. As such, research is warranted to capture students' perceptions across the Chinese mainland and provide a holistic view of NMET washback on their senior high school English learning.

With these research needs in mind, the main purpose of the present study was to investigate—to inve the lens of students from different provincial areas—the short-term NMET washback on their before-test English learning outcomes in senior high school. Specifically, we addressed three research questions:

(1) How do students view the washback of the NMET on their English learning outcomes in senior high school?

(2) How does the background variable of home province influence student perceptions?
(3) What changes do students perceive are needed to make the NMET more useful for senior high school English learning outcomes?

Methods

We designed a survey in the form of a questionnaire to answer our three research questions. The theoretical rationale for conducting survey research is to effectively collect a huge amount of information on attitudes and perceptions from a large group of participants (Mackey & Gass, 2022). Well-constructed questionnaires offer high efficiency and practicality in terms of research time and human, material, and financial resources (Dörnyei, 2010). They are also as much versatile as they are cost-effective since they can be developed to investigate a wide variety of topics with a variety of people in a variety of situations, using specialized item types (Dörnyei, 2010). We found these advantages of questionnaires well-suited for the needs of this large-scale nationwide study.

Research instrument: questionnaire

This research was part of a nationwide project to investigate NMET washback on English teaching and learning in China, for which a questionnaire in Mandarin Chinese was designed and validated. For the present study, we only used the data from relevant questionnaire items in relation to the short-term before-test NMET washback on senior high school English learning. The validation of this questionnaire was conducted in two phases. The first phase involved obtaining qualitative input, and the second phase concerned piloting the questionnaire for statistical analyses.

Phase 1: qualitative input

We drafted the questionnaire on the basis of a review of relevant national policies and government files (including the *NMET test manual* (MoE, 2022), the *Gaokao Assessment System* (NEEA, 2019), the *Curriculum Standards for General Senior High Schools (2017 Edition)* (MoE, 2018), *Suggestions of the State Council of China on Deepening the Reform of Examination and Enrollment System* (State Council of China, 2014), documents on NMET reform measures, schedules, and action plans for individual provincial areas), and an examination of the theoretical and empirical washback studies referred to in the literature review of this article. We then held two consultation meetings with 12 outside experts in language teaching, learning, and assessment, and educational measurement for advice and suggestions on the content, item types, and item grouping and sequencing of the questionnaire. Revisions and alterations to the questionnaire resulted from these two consultations.

Phase 2: piloting

Two pilot studies were conducted for further modifications, validation, and justification of the questionnaire. The initial piloting involved 24 first-year undergraduate students from three Beijing-based universities, who first filled out the questionnaire and then shared their views of and suggestions for it. Revisions to the questionnaire were made according to their advice. In the second pilot study, the revised questionnaire was administered to 200 undergraduates 600 first-year undergraduate students from nine universities (four in Jiangsu province, three in Guangdong, one in Shannxi, and one in Sichuan). Data collection and analysis for this piloting was in every way similar to the procedures for the main study. Based on the results of the second piloting and another consultation with outside experts, minor changes were made to yield the finalized questionnaire.

The final questionnaire

The questionnaire items relevant to NMET washback on senior high school English learning outcomes consisted of three parts (see Appendix 1 for the English version). Part One included 12 multiple choice items about participants' background information. Part 2 comprised two categories. The first category (13 items) concerned participants' perceptions of how important NMET washback should be on their senior high school English learning outcomes, and the second category (13 items) aimed to gather participants' evaluations on actual NMET washback on their senior high school learning outcomes. These items cover the listening, speaking, reading, and writing subskills tested on the NMET that are specified both in the test manual (MoE, 2022) and the *English Curriculum Standards for General Senior High Schools (2017 Edition)* (MoE, 2018). For listening, the tested subskills include understanding simple daily conversations, discussions on general topics, and broadcasts and news. For speaking, the NMET measures the subskills of exchanging factual information, interacting with ease, and speaking in appropriate manners. In the meantime, the NMET reading part involves reading simple factual texts, simple texts in publications, and simplified literature, while the writing part targets candidates' subskills to report information and data, write clear and well-structured texts, and summarize the given texts.

All the 26 items in Part 2 were developed on a 5-point Likert scale of agreement running from "1 = strongly disagree" to "5 = strongly agree". Part 3 contained two open-ended questions—one intended to obtain participants' suggestions on how to make the NMET more useful for senior high school English learning outcomes, and the other to elicit their comment on any issues not covered in previous items.

Participants

Participants in the study were 20,062 first-year undergraduate students who had taken *gaokao* and gained admission to undergraduate programs in 103 higher education institutions across the Chinese mainland. Their data were collected within the first month of their undergraduate study to ensure the trustworthiness of their recollections and accounts concerning the NMET and its washback on their senior high school English learning. Given that only one summer break (approximately 3 months) separated them from their senior high school experience, their memories were less prone to contamination by the elapse of time.

The list of participating universities (19 *985 Project* universities, 19 *211 excluding 985 Project* universities, and 65 ordinary universities[3]) was decided through a synthesis of multistage sampling, stratified sampling, and Probability-Proportional-to-Size sampling. The academic affairs offices of the 103 universities were asked to select 200 first-year undergraduate students from various undergraduate programs to participate in the study. Participants logged on to the designed website of the project to first sign the consent form and then fill out the questionnaire. A total of 20,600 questionnaires were returned, among which 20,062 were valid.

The participants featured diversity of their socio-demographic background and individual characteristics:

- Age: 17,778 (88.62%) were in their late teenage years and 2284 (11.38%) were in their early 20 s;
- Gender: 10,500 (52.34%) are female and 9,562 (47.66%) are male;
- Status of higher education institution: 3114 (15.52%) were studying at *985 Project* universities, 3804 (18.96%) at *211 excluding 985 Project* universities and 13,144 (65.52%) at ordinary universities;
- Undergraduate program: 6506 (32.43%) were in engineering & technology, 3780 (18.84%) in natural sciences, 3407 (16.98%) in economics & management, 2630 (13.11%) in social sciences & humanities, 1775 (8.85%) in arts, music & sports, 1079 (5.38%) in medical science, 588 (2.93%) in law, and 297 (1.48%) in agriculture & forestry;
- Status of senior high school[4]: 10,902 (54.34%) graduated from key senior high schools, 3588 (17.89%) from city-level ordinary senior high schools, and 5572 (27.77%) from county-level ordinary senior high schools;
- NMET result: 2716 (13.54%) gained 90% or more of the total score of the NMET version administered in their provincial regions, 6173 (30.77%) gained 80 to 89.99%, 4779 (23.82%) obtained 70 to 79.99%, 3322 (16.56%) scored 60 to 69.99%, and 3072 (15.31%) gained less than 60%;
- NMET oral test experience: 11,562 took the test while 8500 did not;
- Parents' occupations and highest education levels: see Appendix 2;
- Home provinces/where they received senior high school education (most relevant to this study): spanning across all 31 provinces, autonomous regions, and municipalities in the Chinese mainland (referred to as provinces or provincial areas throughout this paper). Table 6 presents the frequency distribution of participants' home provinces.

[3] Higher education institutions in the Chinese mainland are categorized in large part according to their academic accomplishments. In 1995, 112 universities were selected for the MoE's *211 Project* as the best universities in the Chinese mainland at that time. A second MoE initiative (the *985 Project*) further selected 39 of the 112 *211 Project* universities that have hitherto obtained the largest share of resources and are hence the most sought-after by *gaokao* candidates. Universities not on the list of the *211 Project* are considered ordinary universities, constituting approximately 96% of all higher education institutions in the Chinese mainland.

[4] Senior high schools in the Chinese mainland are classified into key schools—the most eminent, city ordinary schools, and county ordinary schools. Key schools receive the most funds, have more qualified teaching staff, and are therefore more appealing to the best students.

Table 6 Participants' home province

Province	Count	Percentage	Province	Count	Percentage
Jiangsu	1921	9.58%	Zhejiang	385	1.92%
Shandong	1747	8.71%	Shanghai	363	1.81%
Hebei	1713	8.54%	Guizhou	317	1.58%
Henan	1645	8.20%	Fujian	293	1.46%
Liaoning	1377	6.86%	Guangxi	292	1.45%
Guangdong	1136	5.66%	Jiangxi	289	1.44%
Hubei	1091	5.44%	Inner Mongolia	273	1.36%
Shaanxi	1060	5.28%	Heilongjiang	263	1.31%
Anhui	1033	5.15%	Chongqing	252	1.26%
Sichuan	896	4.47%	Xinjiang	224	1.12%
Gansu	746	3.72%	Jilin	211	1.05%
Yunnan	723	3.60%	Tianjin	162	0.81%
Beijing	474	2.36%	Ningxia	148	0.74%
Shanxi	463	2.31%	Hainan	99	0.49%
Hunan	389	1.94%	Qinghai	70	0.35%
			Xizang	7	0.03%
Total				**20,062**	**100%**

Data analysis

We entered the quantitative data from closed-ended items into Stata 16 to generate descriptive and inferential statistics. In order to verify that the Likert scale items were functioning as intended and to identify clusters of items, we conducted an exploratory factor analysis and Cronbach's alpha reliability analysis.

Four rules were applied for factor extraction to ensure meaningful interpretability and practicality: (1) a factor's eigenvalue needed to be greater than 1; (2) any item with a factor loading below 0.50 was deleted and not counted in any extracted factor; (3) a factor needed to have no less than three items; and (4) when double loadings occurred, decisions were made according to meaningful classification. As a result, four factors were extracted and retained, explaining in combination 82.00% of the variance (see Table 7). Cronbach's alphas of the four factors were all above 0.91, indicating high internal consistency and sufficient reliability. The four factors were labeled respectively as NMET washback on speaking, writing, reading, and listening.

A multivariate linear regression analysis was then performed, using ordinary least squares (OLS) method, to examine the strength of relationships between the variable of participants' home provinces and their answers to the 5-point Likert scale items on NMET washback on their senior high school English learning outcomes. The assumptions of OLS regression analysis were assessed for the relevant quantitative data. Scatterplots of the dependent variable against each independent variable showed roughly linear patterns, satisfying the assumption of linearity. The Durbin-Watson test did not find significant autocorrelation in the residuals, revealing independence of errors. Additionally, plots of residuals against predicted values did not show a discernible pattern or a significant change in variance, supporting the assumption of homoscedasticity. Moreover, visual inspection of Q-Q plots suggested that the residuals were largely normally distributed, thereby meeting the assumption of normality.

Table 7 Factor analysis results of the 5-point Likert scale items

Items	Factor				Communality
	1	2	3	4	
18. Speaking: expressing thoughts and feelings (Imp.)	.81				.81
17. Speaking: interacting with ease (Imp.)	.81				.81
31. Speaking: expressing thoughts and feelings (SQ)	.80				.81
30. Speaking: interacting with ease (SQ)	.80				.80
19. Speaking: speaking in appropriate manners (Imp.)	.79				.79
32. Speaking: speaking in appropriate manners (SQ)	.79				.78
16. Speaking: exchanging factual information (Imp.)	.78				.78
29. Speaking: exchanging factual information (SQ)	.77				.75
37. Writing: summarizing texts (SQ)		.79			.80
38. Writing: reporting information and data (SQ)		.79			.80
23. Writing: writing clear and well-structured texts (Imp.)		.79			.79
36. Writing: writing clear and well-structured texts (SQ)		.79			.78
24. Writing: summarizing texts (Imp.)		.77			.76
25. Writing: reporting information and data (Imp.)		.77			.75
20. Reading: reading simple texts in publications (Imp.)			.78		.78
33. Reading: reading simple texts in publications (SQ)			.78		.77
35. Reading: reading simple factual texts (SQ)			.78		.77
22. Reading: reading simple factual texts (Imp.)			.77		.76
21. Reading: reading simplified literature (Imp.)			.77		.76
34. Reading: reading simplified literature (SQ)			.77		.76
15. Listening: understanding broadcasts and news (Imp.)				.78	.77
28. Listening: understanding broadcasts and news (SQ)				.78	.77
14. Listening: understanding discussions on general topics (Imp.)				.77	.75
27. Listening: understanding discussions on general topics (SQ)				.76	.74
13. Listening: understanding simple daily conversations (Imp.)				.72	.67
26. Listening: understanding simple daily conversations (SQ)				.70	.63
Eigenvalue	14.46	7.66	5.89	2.83	
Variance explained (%)	40.38	17.92	15.62	7.42	
Accumulated variance explained	40.38	58.30	73.92	82.00	
Cronbach's alpha	.95	.94	.92	.91	

(1) Extraction Method: Principal Component Analysis; (2) Rotation Method: Varimax with Kaiser Normalization; (3) Factor loadings lower than 0.5 are suppressed in the table; (4) *Imp.* Importance, *SQ* Status Quo

Furthermore, examination of VIF values (all < 10.0) indicated no multicollinearity among the independent variables. The control variables held fixed in the regression analysis were participants' age, gender, status of higher education institution (*985 Project*, *211 excluding 985 Project*, or ordinary), undergraduate program/major, status of senior high school (key, city ordinary, or county ordinary), NMET result, NMET oral test experience, parents' occupations, and parents' highest levels of education.

The analysis of the qualitative data from the two open-ended questions (86, 643 Mandarin Chinese characters) was conducted by entering participants' answers to a keyword analysis, using the Word List function and Concordance tools in AntConc 3.5.0, a corpus search and concordancing program. In so doing, regular and recurrent concordances, collocations, and expressions were identified and extracted. The analysis was carried out in Mandarin Chinese, the same language in which the participants provided their answers. The citations, excerpts, and expressions we refer to in this paper are the English translations of the original data. We conducted the initial translation and then submitted it to an outside expert in Chinese English translation for proofreading. Any losses of meaning or mistranslations were corrected, and a connotation-denotation distinction was made to retain the subtle nuances in participants' answers.

Results

Findings are organized and presented by the three research questions related to (1) student perceptions of NMET washback on senior high school English learning outcomes; (2) the influence of students' home provinces on their perceptions; and (3) their views on the necessary changes to make the NMET more useful for senior high school English learning outcomes.

RQ 1 How do students view the washback of the NMET on their English learning outcomes in senior high school?

To answer this question, we use the descriptive statistics of the 26 five-point Likert scale items (see Appendix 3 for the descriptives) and the supplementary remarks on the open-ended survey questions. Results are presented in correspondence with the four fundamental abilities tested in the NMET—ted in t (the first factor, eight items), writing (the second factor, six items), reading (the third factor, six items), and listening (the fourth factor, six items).

Speaking

The descriptives showed that survey respondents attached great importance to NMET washback on senior high school students' English speaking ability. On a scale of five, all the four relevant items (Items 16–19) received students' high ratings, with the lowest being 4.23 (Item 18). This indicated students' deeply held belief in the strong washback that the NMET should produce on senior high school students' progress in English speaking skills. There was a relatively higher expectation that the NMET and the before-test preparation process would improve students' English skills to exchange factual information through oral communication (Mean = 4.34, S.D. = 0.99) and to speak in appropriate manners (Mean = 4.31, S.D. = 1.03).

Meanwhile, the students' ratings on the status quo of these four issues revealed their general satisfaction with the actual NMET washback on senior high school learning outcomes in English speaking (Items 29–32). Among the four surveyed subskills, the ability to carry out oral exchanges of factual information received the highest mean rating (Item 29, Mean = 3.48, S.D. = 1.14), and the actual NMET washback on the other three aspects also received respondents' positive evaluations. Noteworthy, there was a wide gap between the status quo and students' expectations. This larger than 1-point difference between hope and reality denoted the considerable efforts needed to create stronger and more positive NMET washback on senior high school students' English learning outcomes in speaking.

Writing

Concerning the importance of NMET washback on senior high school students' learning outcomes in English writing, students' mean ratings of the three relevant items (Items 23–25) all exceeded four, indicating a widely shared opinion that the NMET should lead to stakeholders' efforts to improve senior high school students' English writing competence. Among the three items, writing clear and well-structured texts was rated as the most important writing subskill that senior high school students needed to develop for the purpose of overcoming the challenges posed by the NMET (Item 23, Mean = 4.17, S.D. = 1.03). Survey participants' comments on the open-ended questions provided a plausible explanation for the perceived more importance of this subskill than the other two. According to 29 respondents, the ability to write cohesive and coherent texts is the foundation on which students build their competence to write for specific purposes, and therefore in comparison with performing specific summarizing and data reporting tasks, it is the basic skill of writing clear and well-structured texts that requires more attention in senior high school English learning as well as in the NMET preparation process.

As to the actual NMET washback on senior high school students' English writing ability, all three items (Items 36–38) received positive evaluations, showing students' overall satisfaction with the efficacy of actual NMET washback on the development of students' English writing proficiency. Of the three subskills in question, writing clear and well-structured texts was the one reported to have improved the most in senior high school English learning (Item 36, Mean = 3.35, S.D. = 1.18). In view of the above finding that students attached the greatest importance among these three subskills also to writing clear and well-structured texts, their evaluation on its status quo suggested the practical effectiveness of NMET washback in matching students' expectation and addressing their major concern over writing-learning outcomes in senior high school.

Reading

Students considered it important for the NMET to produce positive washback on senior high school students' learning outcomes in English reading—ihe average ratings on the three items (Items 20–22) were all above 4, indicative of the high expectation that the NMET facilitates improvement in senior high school students' English reading ability. Among the three investigated subskills, reading simple factual texts was perceived the most important for students to develop in senior high school English learning (Item 22, Mean = 4.18, S.D. = 1.05).

In the meantime, overall perceptions of actual NMET washback on senior high school students' English reading achievement were positive (Items 33–35). Of the three relevant survey items, reading simple factual texts (Item 33, Mean = 3.36, S.D. = 1.18) and reading simple texts in publications (Item 35, Mean = 3.32, S.D. = 1.19) were the two subskills students reported that they had made greater improvements in senior high school English learning. On the other hand, students' ability to read simplified English literary works was relatively underdeveloped (Item 34, Mean = 3.29, S.D. = 1.20). This finding manifests the powerful washback that the NMET has exerted on senior high school students' English reading development. The reading passages in the NMET have mainly been selected or adapted from factual texts published in books, newspapers, and journals; yet seldom do they relate to literary genres—even though they are listed in the NMET test manual as an option for reading material selection, hence survey respondents' less self-stated satisfaction with their English learning outcomes in this subskill.

Listening

Student responses to the three relevant items (Items 13–15) revealed their perception that the NMET should be the driving force to promote senior high school students' English listening ability. Students felt that understanding simple interactions about daily life (Item 13, Mean = 4.39, S.D. = 0.98) and understanding discussions/talks on general topics (Item 14, Mean = 4.34, S.D. = 0.98) were relatively more important of the three surveyed subskills that needed students' direct efforts and involvements in senior high school English learning.

Comparatively, survey respondents held that the NMET washback on senior high school students' ability to understand broadcasting and news programs did not necessarily need to be as strong as that on the other two investigated skills above. A participant left her remark explaining why she had given a low rating on this item:

> *"Actually, I cannot recall any previous NMET listening subtest that asked candidates to listen to radio broadcasts or live news and answer relevant questions. I guess maybe the NMET does not intend to test students' performance in that area, so I think its effects on senior high school English learning is not and should not be very strong" (Respondent No. 747).*

This comment revealed that the text types of the materials selected for the NMET listening subtest contributed to student ratings on the three items. The NMET listening subtest, irrespective of its various forms, has indeed been using mostly daily conversations or discussions on general topics as the listening materials; therefore, it is understandable that Item 13 (understanding simple interactions about daily life) and Item 14 (understanding discussions/talks on general topics) obtained higher average ratings due to the strong and immediate needs for senior high school students to enhance their development in these two aspects to perform well on the NMET listening subtest.

Differences were also found between student perceptions of the actual NMET washback on these three listening subskills (Items 26–28). While in general survey respondents held positive attitudes toward the actual NMET washback on senior high school students' listening development, their abilities to understand simple conversations about daily life (Item 26, Mean = 3.59, S.D. = 1.12) and to understand discussions or talks on general topics (Item 27, Mean = 3.48, S.D. = 1.12) were perceived to have made more gains in their senior high school English learning. It is of no surprise to see greater NMET washback on students' English listening learning attainment in these two aspects given that the NMET listening texts mainly pertain to conversations set in an everyday social context and dialogues or monologues related to general subject matters.

To sum up, survey respondents expressed high hopes for strong and positive NMET washback to enhance senior high school students' English learning outcomes and gave positive evaluations of the actual washback on students' development in the four skills assessed in the NMET. However, there were noticeable differences between participants'

Table 8 Estimated significant effects of home province on the importance of NMET washback on English speaking-learning outcomes

	Dependent variable			
	Item 16	**Item 17**	**Item 18**	**Item 19**
Home province[a]				
Beijing	0.165***	0.170***	0.208***	0.175***
	(15.05)	(4.97)	(6.53)	(9.69)
Guangdong	0.149***	0.163***	0.184***	0.109***
	(5.95)	(6.28)	(6.56)	(4.25)
Jiangsu	0.137***	0.159***	0.158***	0.116***
	(7.12)	(8.95)	(7.63)	(7.66)
Shanghai	0.110***	0.106***	0.116***	0.121***
	(5.22)	(5.47)	(7.30)	(6.88)
Zhejiang	0.148***	0.116***	0.119***	0.130***
	(5.69)	(7.74)	(5.64)	(8.87)
Ctrl variables	Yes	Yes	Yes	Yes
N	20,062	20,062	20,062	20,062
R-squared	0.073	0.067	0.071	0.068

NOTE: *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; *t* statistics in parentheses; [a] The omitted category is Gansu province

expectations and realities. Among the four skills, students' English speaking ability needed the greatest sustained efforts to close the gap.

RQ 2 How does the background variable of home province influence student perceptions?

To address this question, we use findings from the OLS regression analysis and student comments on the open-ended survey questions. Results are presented in the same order as that for RQ 1—speaking, writing, reading, and listening.

Speaking

Regression analyses revealed that students' home province was a strong predictor of their views in relation to NMET washback on senior high school English speaking-learning outcomes. When the control variables were held constant, survey respondents originally from Beijing, Guangdong, Jiangsu, Shanghai, and Zhejiang were associated with significantly higher expectations compared to their countrymen from Gansu (the arbitrarily selected control group) for the NMET to produce strong and positive washback on senior high school students' English speaking development (see Table 8, $p < 0.001$ for all comparisons). This finding showed the greater support from students originally from these five provincial areas for higher NMET guiding power to improve senior high school students' English speaking competence.

Meanwhile, the OLS regression analyses found that students taking the NMET in these five regions were also linked to significantly higher ratings than those from Gansu to the four surveyed items pertaining to actual NMET washback on senior high school students' English speaking achievement (see Table 9, $p < 0.001$ for all comparisons). This finding suggested that stakeholders from these provinces and municipalities had

Table 9 Estimated significant effects of home province on the actual NMET washback on English speaking-learning outcomes

	Dependent variable			
	Item 29	**Item 30**	**Item 31**	**Item 32**
Home province[a]				
Beijing	0.155^{***}	0.166^{***}	0.152^{***}	0.158^{***}
	(4.23)	(4.29)	(4.25)	(4.18)
Guangdong	0.142^{***}	0.142^{***}	0.144^{***}	0.164^{***}
	(4.14)	(4.15)	(4.09)	(5.00)
Jiangsu	0.052^{***}	0.034^{***}	0.043^{***}	0.030^{***}
	(3.47)	(1.29)	(2.93)	(2.00)
Shanghai	0.147^{***}	0.151^{***}	0.161^{***}	0.155^{***}
	(4.19)	(4.23)	(4.28)	(4.25)
Zhejiang	0.041^{***}	0.028^{***}	0.067^{****}	0.043^{***}
	(2.04)	(1.04)	(3.58)	(2.34)
Ctrl variables	Yes	Yes	Yes	Yes
N	20,062	20,062	20,062	20,062
R-squared	0.027	0.020	0.020	0.019

NOTE: *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; *t* statistics in parentheses; [a] The omitted category is Gansu province

invested greater efforts to enhance students' English speaking skills in senior high school education and gained effectiveness.

The responses from Beijing, Guangdong, and Shanghai students were noteworthy in yet another sense. At the time of writing, these three regions are the only provincial areas that require all their NMET test takers to sit for a listening-speaking integrated component. It would thus come as no surprise if survey respondents from these areas were among those whose English speaking skills made greater progress in senior high school English learning. As demonstrated in Table 9, NMET test takers from Beijing, Guangdong, and Shanghai indeed gave significantly higher evaluation to the NMET's actual washback on the development of senior high school students' English oral production ability (i.e., the skills tested in the NMET receive focal attention and intensive training, and in turn lead to eye-catching gains in students' learning outcomes).

Writing

Respondents' home province was significantly related to their perceptions toward NMET washback on senior high school students' English writing development. With respect to expectations, although no significant inter-province difference was found in student views about the importance of NMET washback on their ability to write clear and well-structured texts (Item 23, $p > 0.05$ for all comparisons), students originally from Beijing, Guangdong, and Shanghai held significantly higher hope than their Gansu counterparts for greater competence in text summarizing and data reporting to meet NMET demands on test takers' English writing performance (see Table 10, Items 24 and 25, $p < 0.001$ for all comparisons).

Regarding the actual NMET washback, no significant relationship was observed between students' province of origin and their ratings on the effectiveness of NMET

Table 10 Estimated effects of home province on the importance of NMET washback on English writing-learning outcomes

	Dependent variable		
	Item 23	**Item 24**	**Item 25**
Home province[a]			
Beijing	0.008	0.045***	0.060***
	(0.19)	(5.97)	(6.01)
Guangdong	0.006	0.071***	0.043***
	(0.11)	(7.41)	(5.19)
Shanghai	0.005	0.063***	0.051***
	(0.16)	(6.99)	(5.82)
Ctrl variables	Yes	Yes	Yes
N	20,062	20,062	20,062
R-squared	0.066	0.056	0.054

NOTE: $^*p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$; *t* statistics in parentheses; [a] The omitted category is Gansu province.

Table 11 Estimated effects of home province on the actual NMET washback on English writing-learning outcomes

	Dependent variable		
	Item 36	**Item 37**	**Item 38**
Home province[a]			
Beijing	0.008	0.069***	0.047***
	(0.25)	(2.25)	(3.42)
Guangdong	0.036	0.074***	0.066***
	(1.84)	(2.63)	(3.81)
Shanghai	0.008	0.052***	0.060***
	(0.25)	(2.54)	(3.96)
Ctrl variables	Yes	Yes	Yes
N	20,062	20,062	20,062
R-squared	0.025	0.022	0.021

NOTE: $^*p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$; *t* statistics in parentheses; a The omitted category is Gansu province

washback to enhance senior high school students' ability to write clear and well-structured English texts (Item 23, $p > 0.05$ for all comparisons). A possible explanation is that all the writing subtests of the several NMET forms assess test takers' English production skill to write logically cohesive and structurally coherent texts, and thus senior school students from different provincial areas across the Chinese mainland may have all received sufficient practice in this facet of English writing, hence the insignificant inter-provincial difference. On the other hand, respondents originally from Beijing, Guangdong, and Shanghai were relative to significantly higher ratings to the actual NMET washback on their development of text-summarizing and data reporting skills (see Table 11, Items 37 and 38, $p < 0.001$ for all comparisons).

Table 12 Estimated effects of home province on the actual NMET washback on English reading-learning outcomes

	Dependent variable		
	Item 33	**Item 34**	**Item 35**
Home province[a]			
Beijing	0.039	0.203***	0.009
	(0.86)	(8.34)	(0.47)
Guangdong	0.050	0.155***	0.002
	(0.78)	(6.25)	(0.06)
Shanghai	0.024	0.315***	0.048
	(0.81)	(9.81)	(1.51)
Control variables[c]	Yes	Yes	Yes
N	20,062	20,062	20,062
R-squared	0.058	0.057	0.052

NOTE: *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; *t* statistics in parentheses; [a] The omitted category is Gansu province

Reading

The OLS regression analysis revealed that the variable of home province was not predictive of student perceptions with respect to the importance of NMET washback on senior high school students' competence in English reading comprehension (Items 20–22, $p > 0.05$ for all comparisons). This finding is a likely result considering that the design of the reading test and the range of skills examined in this component were fairly consistent across the NMET forms. Survey respondents from different regions might therefore have a commonly shared view about the active role that the NMET should play in facilitating senior school students' English reading development.

Similarly, home province was also found to have no significant effect on student evaluations in relation to the actual NMET washback on senior high school students' ability to understand simple texts in publications (Item 33, $p > 0.05$ for all comparisons) and to read simple factual texts (Item 35, $p > 0.05$ for all comparisons)—two skills specifically assessed in the NMET reading test. This finding suggests a wide agreement among test takers concerning the effectiveness of NMET washback in promoting senior high school students' English learning outcomes in the NMET-targeted reading skills. However, as regards the non-NMET-targeted skill to read simplified literary works (Item 34), survey respondents from Beijing, Guangdong, and Shanghai were associated with significantly greater satisfaction with their development in this aspect during senior high school (see Table 12, $p < 0.001$ for all comparisons).

Listening

Results showed that home province was a significant predictor variable of student ratings on the importance of NMET washback on English listening-learning outcomes in senior high school. Specifically, after controlling for other variables of interest, respondents originally from Heilongjiang, Henan, Liaoning, and Shanxi were correlated with significantly higher ratings on the three relevant survey items in this respect (see Table 13, $p < 0.001$ for all comparisons). It is noteworthy that all these four provinces were among

Table 13 Estimated significant effects of home province on the importance of NMET washback on English listening-learning outcomes

	Dependent variable		
	Item 13	**Item 14**	**Item 15**
Home province[a]			
Heilongjiang	0.065***	0.076***	0.083***
	(2.56)	(2.61)	(2.67)
Henan	0.069***	0.075***	0.072***
	(1.98)	(2.22)	(2.17)
Liaoning	0.063***	0.072***	0.088***
	(3.60)	(3.28)	(2.88)
Shanxi	0.074***	0.087***	0.071***
	(5.99)	(2.13)	(2.72)
NMET listening test: No[b]	0.045***	0.053***	0.047***
	(2.94)	(2.89)	(2.82)
Ctrl variables	Yes	Yes	Yes
N	20,062	20,062	20,062
R-squared	0.074	0.073	0.060

NOTE: *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; *t* statistics in parentheses; [a] The omitted category is Guangdong province; [b] The omitted category is NMET listening test: Yes

the seven provincial areas that, as of the year 2022, did not count test takers' listening scores in their total NMET results or did not even administer the NMET listening subtest. In fact, when survey respondents from all these seven provincial regions were held as one subgroup, they were observed to be significantly related to higher ratings on the three items than their countrymen from provincial areas where the NMET listening component was mandatory (also see Table 13, $p < 0.001$ for all comparisons).

Two respondents' comments might help explain this finding:

> *"We were not asked to take the NMET listening subtest, but I don't think that will do us any good in the long run. I strongly hope the listening subtest can be added to the NMET in my province so that stakeholders will care about students' English listening competence. You know, no test, no attention" (Respondent No. 4159, home province: Shanxi).*
>
> *"I don't understand why Henan does not administer the NMET listening subtest. I think we certainly have the facilities and equipment to do so. Some would say we are lucky since we don't need to prepare for the listening subtest, but we are really not!!! A lot of teachers, students, and even parents want the NMET listening subtest because this would lead to practice and improvement of students' listening skills in senior high school" (Respondent No. 140, home province: Henan).*

These remarks indicated that the exclusion of the NMET listening subtest was not a welcomed decision. NMET test takers from relevant provinces were aware of the power that the NMET had on senior high school English teaching and learning—the intensity of NMET washback, and it was in their belief that the NMET should include a listening part to produce pedagogical changes to senior high school English education and direct students' learning efforts to improve their English listening proficiency. It might be

Table 14 Estimated significant effects of home province on the actual NMET washback on English listening-learning outcomes

	Dependent variable		
	Item 26	**Item 27**	**Item 28**
Home province[a]			
Gansu	− 0.033***	− 0.038***	− 0.030***
	(− 1.61)	(− 1.76)	(− 1.74)
Heilongjiang	− 0.059***	− 0.055***	− 0.064***
	(− 2.81)	(− 2.36)	(− 2.41)
Henan	− 0.041***	− 0.033***	− 0.038***
	(2.33)	(− 1.93)	(− 1.51)
Liaoning	− 0.040***	− 0.041***	− 0.053***
	(− 2.77)	(− 2.26)	(− 2.51)
Shanxi	− 0.048***	− 0.040***	− 0.030***
	(− 1.84)	(− 2.08)	(− 1.42)
Xinjiang	− 0.060***	− 0.055***	− 0.042***
	(− 2.92)	(− 1.61)	(− 1.78)
NMET listening test: No[b]	− 0.132***	− 0.147***	− 0.166***
	(− 4.02)	(− 4.53)	(− 6.85)
Ctrl variables	Yes	Yes	Yes
N	20,062	20,062	20,062
R-squared	0.026	0.028	0.021

NOTE: *p < 0.05, ** p < 0.01, *** p < 0.001; *t* statistics in parentheses; [a] The omitted category is Guangdong province; [b] The omitted category is NMET listening test: Yes

owing to this expectation for strong and positive NMET washback that survey respondents originally from provincial areas with no required NMET listening subtest assigned significantly greater importance to potential NMET washback on senior high school students' learning outcomes in English listening comprehension.

Meanwhile, the regression analysis found that the variable of home province was also significantly predictive of student views about the actual NMET washback on senior high school students' English listening development. When control variables were held constant, ratings given by respondents from Gansu, Heilongjiang, Henan, Liaoning, Shanxi, and Xinjiang were significantly lower on the effectiveness of actual NMET washback in advancing senior high school students' English listening skills (see Table 14, $p < 0.001$ for all comparisons).

All these six provincial areas were among the seven regions in the year 2022 that did not hold a mandatory NMET listening subtest. When survey respondents from all the seven regions were treated as one group that did not sit for a compulsory NMET listening component, the regression analysis discovered that this group was highly correlated with a decrease in reported satisfaction with the current state of NMET washback in terms of producing anticipated English learning outcomes in the three surveyed listening skills for senior high school students (also see Table 14, $p < 0.001$ for all comparisons).

Thus, overall, the variable of home province was significantly correlated with and predictive of the respondents' views about the importance and status quo of NMET washback on senior high school students' English speaking, writing, and listening

Table 15 Concordance frequencies for the open-ended survey question No. 39 (Top 10)

Concordance	Frequency	Group 1	Group 2
A mandatory listening test	774	✓	
A mandatory speaking test	727	✓	
One uniform nationwide NMET form	681		✓
One NMET form across all provincial areas	648		✓
All-round development of English ability	626	✓	
A mandatory listening and speaking test	612	✓	
Same requirements for all provincial area	587		✓
Conducive to filling inter-provincial gaps in the quality of senior high school English education	551		✓
Conducive to solving the problem of inter-provincial unfairness in college enrollment	535		✓
Educational equality for all NMET test takers regardless of their home province	460		✓

development, despite the varying strengths of the relationships. On the other hand, no significant difference was found between respondents from different provincial areas in their perceptions toward NMET washback on students' English reading learning outcomes in senior high school.

RQ 3 What changes do students perceive are needed to make the NMET more useful for senior high school English learning outcomes?

Survey respondents' answers to the open-ended Question 39 and part of their comment on Question 40 in the questionnaire concerned their suggestions on necessary changes for the NMET to further enhance the outcomes of senior high school English learning. Their answers were consistent with their expectations and perceived reality of NMET washback, and with the differences in perception between students from different home provinces. The keyword analysis for the open-ended item yielded a list of the ten most recurrent concordances in respondents' propositions that could be classified into two categories: Category 1—to administer mandatory speaking and/or listening subtests to all NMET test takers, and Category 2—to use one nationally unified NMET form across all provincial areas (see Table 15).

Making the listening subtest mandatory was the most frequently proposed change to the NMET, all by survey respondents from the provincial areas delivering listening-excluded NMET forms. These participants enunciated their high hopes that their home provinces would adopt the NMET listening test as soon as possible. It was in their belief that this change to the NMET in their home provinces would lead to strong and desirable washback on senior high school English learning because all stakeholders would subsequently devote extensive attention and sufficient effort to improve students' English listening comprehension ability. In addition, survey respondents felt that since English learning in tertiary education involved a variety of listening tasks and activities, earlier and prolonged engagement with English listening learning and instruction in senior high school would contribute to adequate preparation for students to manage college requirements.

This rationale also applies to survey respondents' suggestion for a compulsory NMET speaking subtest, put forward by both students who took this subtest and those who did not. Respondents stated that their English speaking competence was relatively underdeveloped compared to other aspects of language, which they attributed mainly to the optional nature of the NMET speaking subtest in most provincial regions. They reported that many stakeholders they knew were cognizant of the potential beneficial washback that would be produced by a mandatory NMET speaking component on senior high school students' English speaking-learning outcomes—and in turn on college English learning, but pitifully the status quo was not to students' advantage in this regard.

Many survey respondents recognized that it was particularly difficult to administer a nationwide mandatory NMET speaking subtest given the existing wide inter-provincial variation across the Chinese mainland, especially in terms of available human, equipment, and facility resources. Students and other stakeholders therefore tried their personal best to enhance students' English speaking development in their own ways (e.g., schools offering specialized speaking courses, teachers bringing to students' attention the importance of English speaking skills to their all-round development and potential future studies to circumvent the weaker than expected NMET washback on senior high school English learning outcomes in spoken English production. The optional nature of the NMET oral subtest and hence lack of anticipated washback was perceived to have marginalized English speaking learning and instruction in senior high school, and the bottom-up endeavors through students, teachers, schools, and parents alone were far from enough to change the situation. Research participants' solution therefore was to make the NMET speaking test mandatory.

The other category of recurring suggestions represented survey respondents' hope for a nationwide uniform NMET form administered across all provincial regions. This proposal was raised by students from not only provinces adopting NEEA-constructed NMET forms but also regions administering their self-developed province-wide forms. According to the former group of respondents, the NEEA-developed NMET forms were of high test qualities, so much so that any one of these forms was qualified to be used across all provincial areas. Additionally, since NEEA is affiliated with the Ministry of Education, a nationwide administration of its NMET form would manifest nationally unified instructional requirements and evaluation criteria for senior high school English education. The NMET results of test takers from different provinces could therefore be straightforwardly compared, through which between-province differences could be revealed and underlying problems identified. Central authorities and policymakers can then provide informed guidance, tailored assistance, and proper accommodation to fill inter-provincial gaps and promote educational equality pertaining to development opportunities and prospects.

A number of students fully understood the practicality and fairness concerns behind the decades-long policy of administering province- rather than nationwide NMET tests, yet it was in their opinion that cross-province differences in educational funding and resources across the Chinese mainland had decreased over the past decade thanks to national-level financial regulation and national distribution. The reasoning of the policy therefore no longer applies, and the time is ripe for a nationally uniform NMET form.

Hundreds of these respondents went even further to propose a reform on the country's college admissions system. Currently, universities in the Chinese mainland have an enrollment quota for each provincial area, meaning students taking the NMET in different regions do not compete with each other for places at higher education institutions. This province-specific quota system has long been criticized for the unfairness it entails since students from more economically developed regions (e.g., Beijing and Shanghai) are accepted to prestigious universities at higher rates than their counterparts from more rural and remote provinces. One respondent wrote:

> *"It's extremely hard to get into top universities if you take the gaokao in our province. The competition is much fiercer than Beijing where many well prestigious universities are located. These universities always reserve more places for students with Beijing hukou, so the lowest qualifying score for a Beijing-based gaokao candidate is much lower than the score required from a student in my province." (Respondent No. 14082, home province: Shandong).*

These regional inequalities in the college admissions system and in the distribution of scarce, quality education resources, and opportunities led to the respondents' call for a termination of the province-specific quota system to address allegations of regional bias. To these respondents, universities should pool together their yearly enrollment quotas which are up for grabs to all *gaokao* candidates irrespective of their home provinces, and administering a nationwide uniform form for the NMET as well as for the other *gaokao* subjects is the critical first step in that direction.

On the other hand, survey respondents from provincial areas administering self-developed NMET forms asked for a nationally unified test form out of a different concern—rumors about the test difficulty of the self-developed NMET forms. As a participant explained his upset:

> *"Many people say the self-developed NMET forms are easier for students to get high scores than the NEEA-constructed ones, but I've never seen them provide any evidence. I don't even think they know what our form looks like. This is so unfair. I myself tried the NEEA-constructed forms for previous years when I was preparing for the gaokao, and I think our self-developed form is just as difficult as the NEEA-developed ones, if not more so." (Respondent No. 7855, home province: Shanghai).*

This frustration was echoed in many other participants' comments. They felt offended by the groundless accusation that the self-developed NMET forms in their home provinces were at a lower level of test difficulty, especially when no official data has confirmed this. These claims were perceived depressing as these students taking self-developed NMET forms thought they had "survived" an NMET preparation just as arduous as that for their countrymen in other provincial areas. It is therefore understandable that these respondents expressed their hope for a nationwide uniform NMET form, because there will no longer be any problem about unfairness in test difficulty and hence the so-called local protectionism in this respect.

In summary, survey respondents called for mandatory NMET listening and speaking subtests in all provincial areas in order to produce strong and positive washback on senior higher school students' English learning achievements across the Chinese mainland. Moreover, participants from provinces adopting the NEEA-constructed forms and from those using their self-developed ones both suggested administering a national-level uniform NMET form across all provincial regions, though these two groups proposed this change on different grounds.

Discussion and conclusion

The NMET serves multiple roles in the Chinese mainland, more than a college admissions exam as manifested in its official name. It is not only a gatekeeper to higher education that could decide students' future prospects; it also relates to students' growth and development through pre-college educational programs and interventions as well as the lives of teachers, parents, policymakers, and other stakeholders. Its high stakes on secondary English education, particularly on curricular agenda and pedagogical practice, warrant constant investigations into stakeholders' expectations for and the actual effectiveness of before-test NMET washback on senior high school English learning.

Two decades ago, Qi (2004) found that the high-stake nature of the NMET had undermined its potential to facilitate senior high school English teaching and learning, and its actual washback did not match test constructors' expectations. This was because the pursuit for high NMET scores had more often than not outweighed the MoE's original intention for the test to direct stakeholders' attention to the development of students' language use abilities, which in some cases even led to sheer teaching-to-the-NMET classrooms. Based on this finding, Qi (2004) suggested that the stakes of the NMET be reduced so as to release senior high school English teachers from the considerable pressure induced by NMET preparation and they thus can invest efforts more toward effecting beneficial change to students' communicative English skills.

However, test stakes are determined by test purposes. Since the principal use of the NMET is to screen in and out candidates for tertiary education, it inevitably entails high stakes, and it is in turn this built-in signature high-stake quality that makes it possible for the NMET to produce strong washback on senior high school English learning. Notice that Qi's (2004) suggestion to lower the NMET's stakes for beneficial washback is justifiable as her major research finding was that the NMET had prompted grueling and lengthy teaching-to-the-test practices in senior high school. On the other hand, the present study found that former NMET test takers were broadly satisfied with the effectiveness of the test's actual washback in enhancing their senior high school English learning outcomes in all four communicative abilities. Even though their competence in the NMET-targeted subskills gained more momentum than those not explicitly assessed in the NMET, the development of the latter also received positive evaluations from test takers, showing a change of situation since Qi's (2004) research that senior high school English learning and instruction now tend to increase, recognize, and support stakeholders' endeavors to promote students' all-round English proficiency, be the specific skills and components tested or not in the NMET.

Having said that, still noticeable is the gap between student expectations and realities. As the findings of this study showed, despite the prevailing positive evaluations in relation to the actual NMET washback on senior high school students' English learning outcomes in all the subskills in question, the status quo trailed behind research participants' expectations in all these aspects. Specifically, the NMET was more effective in facilitating students' reading and writing learning attainment than in listening and speaking, demonstrating the distinct intensity of before-test NMET washback on the development of test takers' different English abilities. This imbalanced competence extends to and eventually influences students' college English learning, regarded as a manifestation of the NMET's long-term washback (Zhang & Bournot-trites, 2021). In Zhang and Bournot-trites' (2021) study on this after-test NMET washback, college students reported that their English speaking and listening skills were not adequate for English college requirements in comparison to their reading and writing capabilities. This finding revealed the inseparability of the short-term and long-term NMET washback and highlighted all the more important work that is needed to fully harness the test's potential to advance all aspects of students' English language development in senior high school.

To achieve this, we suggest fine-tuning the NMET test design. As Bachman and Palmer (2010) have argued, before deciding on a specific test design, test developers should consider the beneficial consequences they intend to bring about and plan what to do to produce these consequences. The intended washback ought to be reflected in the measured abilities and in the assessment tasks developed to elicit test takers' performance on which their competence will be inferred. Also taking this line of reasoning, the NEEA constructed the *Gaokao Assessment Framework* (NEEA, 2019) that charts a course for further reform to deliver quality secondary education leading to effective learning outcomes, and directed by this framework, the NMET committee issued guidelines set out to clarify the orientation and principles for the test's content reform (Chen et al. 2019). However, these guidelines did not specify details about future changes. Given the high stakes of the NMET, it is advisable to hear stakeholders' voice on necessary and practical reform measures.

In this regard, NMET test takers—MET primary stakeholder group—on the present study strongly recommended administering mandatory NMET listening and speaking subtests across all provincial areas in the Chinese mainland, in hopes for attention to and development of senior high school students' English listening and speaking proficiency. Since tests influence learners' learning processes and products (Bailey, 1996), changes to the test design including (but not limited to) test content, format, and delivery mode of the high-stake NMET will wield strong power on the direction (Green, 2003, 2006, 2007) and value (Watanabe, 1997, 2004a) of the test's washback.

A comparison between our findings and those of Qi (2004) and Zhang and Bournot-trites (2021) may verify the potential and suitability of students' proposal for non-optional NMET listening and speaking components. Qi's (2004) research mainly focused on the NMET washback situation in the province of Guangdong and found senior high school teachers' complete disregard for students' English speaking development simply because the NMET did not test this skill back then. When Zhang and Bournot-trites (2021) collected their research data, Guangdong had become the only provincial area in the Chinese mainland requiring all their NMET test candidates to sit a speaking subtest. As their findings showed, student participants originally from Guangdong were significantly linked to higher evaluations on the effectiveness of NMET washback in equipping them with sufficient English speaking competence to handle college workload and facilitate subsequent college English learning, which was attributed by an interviewee from Guangdong to the non-optional nature of the NMET speaking subtest in his province. Furthermore, at the time of the present study, Beijing and Shanghai had joined Guangdong to deliver a compulsory oral subtest to all its NMET test takers, and students from these three regions were found significantly correlated with higher levels of satisfaction with the status quo of NMET washback on senior high school students' English speaking learning outcomes, the scale of which the ratings given by their counterparts from other provincial areas were far behind to catch up with. This gradual and promising change of situation lends support to our research participants' suggestion to hold mandatory NMET speaking and listening subtests nationwide.

The dynamic and evolving landscape of NMET washback in Guangdong foregrounds another focal consideration of this study: the influence of home province on student perceptions toward NMET washback. One of the influential washback hypotheses is that tests have different types and amounts of washback on some learners than on others (Alder & Hamp-Lyons, 1996). This means that variations in learners' individual differences may mediate the relationship between a test and its washback on learning outcomes, and thus lead to subgroup differences in perception. As the present study found, despite the state-level overall positive NMET washback on senior high school students' English learning outcomes, significant inter-province differences in test takers' views were identified in the following two senses.

First, test takers were generally parted into two home province groups in terms of their perceptions toward the actual NMET washback on English listening and speaking learning outcomes in senior high school, depending on whether the skills were tested in their NMET. At the time of writing, test candidates' English listening proficiency was still not assessed in seven provincial areas, and only three regions administer a mandatory speaking section. As a result, this different operationalization of the NMET listening and speaking subtests was associated with significant inter-provincial differences in test takers' views about the NMET's effectiveness in enhancing senior high school students' English listening and speaking development.

Previous research revealed that test takers whose home province delivered non-compulsory NMET listening subtest reported significantly fewer listening activities in their senior high school English learning and instruction (Zhang, 2019). A possible consequence is that students' listening ability would be underdeveloped in their before-NMET English learning, which was held responsible by quite a few NMET test takers on the present study for their significantly lower levels of satisfaction with the NMET's actual washback on their English listening development in senior high school. This may also account for why students originally from these provinces faced great challenges and found themselves with a substantial disadvantage in undergraduate academic study, a negative long-term NMET washback as prior research unveiled (Zhang & Bournot-trites, 2021). Relatedly, the qualitative data of the present study indicated that it was mostly test takers from the provincial areas with no mandatory NMET listening component that called for an inclusion of the subtest in order for strong and facilitative washback on senior high school students' English listening attainment. The case for the NMET speaking subtest is similar. Since Beijing, Guangdong, and Shanghai NMET test takers on this study were all required to take the NMET oral test, it is unsurprising to find that they gave significantly higher evaluations of the status quo of before-test NMET washback on students' English speaking development. Furthermore, students from home provinces having only an optional NMET speaking section were more vocal for the subtest to be made mandatory. These results showed the overpowering washback wielded by test content on students' English learning outcomes, echoing what was reported in washback research on other English tests in the Chinese mainland (e.g., Gu, 2007; Jin, 2000; Zou & Xu, 2017).

Moreover, it also comes to the forefront that relatively more developed provincial regions are pioneers in expediting the shift from "teaching/learning to the NMET" practice in senior high school to all-round development of students' comprehensive English skills. These areas, represented by Beijing, Guangdong, Jiangsu, Shanghai and Zhejiang, feature relatively higher degrees of opening up, economic growth, educational services, and international cultural exchanges. Zhang's (2019) findings pointed to the significantly higher likelihood of students from these regions to learn English for better future job prospects rather than simply for educational opportunities. Relevant participants on Zhang's (2019) study ascribed this result to their extensive hands-on experience of the practical role played by the English language in their lives. This recognition of what a high English proficiency works for their future success might also explain the finding of the present study as to why test takers from these provinces attached significantly greater importance to the NMET's positive washback on the comprehensive development of students' English skills. This demand was best manifested in their perceptions toward the necessity of NMET washback effects on students' speaking learning outcomes in senior high school. Given their deep desire for advancement in their all-round English proficiency and future achievement, these respondents from more developed regions were more vociferous proponents of the NMET's guiding role in leading senior high school English learning efforts to promote students' speaking competence, even though it was not compulsorily tested in the test as is the case with Jiangsu and Zhejiang provinces.

Furthermore, as this study indicated, it was also NMET test takers from these provinces—t w only the NMET-speaking-subtest-mandatory Beijing, Guangdong, and Shanghai—that were associated with significantly higher level of satisfaction with the NMET's actual washback on their English speaking learning outcomes in senior high school. Alongside Zhang's (2019) finding that students from these areas reported a stronger presence of teaching and learning activities related to non-NMET-tested skills in their senior high school English classrooms, we can safely say that relatively more developed provincial areas in the Chinese mainland have been proactive vanguards in shaping senior high school English teaching, learning, and assessment beyond NMET-oriented practice and engagement.

The links this study discovered between before-test NMET washback and the contextual variable of home province are consistent with Shih's (2006, 2007) theoretical implications that extrinsic, intrinsic, and test factors may operate in tandem to influence a test's washback on student learning. In the case of the current study, though home province constitutes a surface-level static factor in and of itself, the underlying forces responsible for the inter-province differences might include extrinsic factors (e.g., provincial *gaokao* policies, socioeconomic circumstances, school contexts, students' future educational and career pathways), intrinsic factors (e.g., students' personal perceptions toward the NMET), and test factors (e.g., the stakes, test content, test format, tested skills, and test purposes of the NMET). These factors might all be at work and jointly affect senior high school students' English learning outcomes. It is a must for policymakers and educators to take all these factors into consideration when formulating plans to produce more beneficial NMET washback and help stakeholders gain deep insights into the dynamic and situated nature of the washback mechanism.

The findings of this study capture a complicated and multifaceted picture of the nuanced NMET washback on students' English learning outcomes in senior high school across different provincial regions, and comprehensive measures as well as cooperative endeavors are particularly called for to render the NMET and its washback maximally beneficial to reduce unfavorable regional disparities in educational affordances and ensure that English learners can take advantage on equal footing of the uniformly standardized yet contextually heterogenous secondary English education across the Chinese mainland. To this end, the MoE has prioritized its agenda to deepen the reform of China's examination and enrollment system (State Council of China, 2014), and correspondingly, the *Gaokao Assessment Framework* (NEEA, 2019) spells out a value-instructed, literacy-oriented, ability-dominant, and knowledge-based assessment framework (NEEA, 2019). This framework specifies the direction, focus, and requirements of the content reform of the *gaokao*, which can promote the systemic test development of the *gaokao* battery, of which the NMET is part, enhance quality education, and advance education equity.

Still within the scope of fairness between provincial areas, a number of the participants on this research suggested administering one same NMET form across the nation to (1) reflect nationally unified standards and requirements of senior high school English teaching, learning, and assessment, identify inter-provincial differences, and subsequently address underlying problems; and (2) serve as a first move toward a national-level quota-pooling system in lieu of the current provincial-specific quota system for college enrollment. Consistent with the participants' first intention, the *China's Standards of English Language Ability* ([CSE], MoE & State Language Affairs Commission, 2018) has been developed and implemented in all stages of Chinese English education. As a unified and scaled framework of English proficiency, the use-oriented CSE, featuring detailed descriptions of specific language use purposes and typical activities, is conducive for a NMET-bridged connection and coherence between senior high school and college English education. As noted earlier, the NMET washback does not only permeate senior high school English learning, it also extends its influence over students' English learning in higher education institutions; therefore, an adoption of the CSE as theoretical reference and content design guidance for NMET development can presumably facilitate not only the intended short-term before-test NMET washback on secondary English learning but also beneficial long-term after-test washback on tertiary English education, as desired by respondents on this study.

It is worth pointing out that that participants' expectation for a nationwide uniform NMET form does not mean there is any problem with the test qualities of the current forms, be they constructed by the NEEA or provincial education departments. The MoE's policy of designating certain provincial areas to develop and administer *gaokao* test forms on their own terms was originally intended to reduce test bias and accommodate the *gaokao* to local contexts and regional milieu, yet Luo and Xiao (2018), taking the NMET as an example, disapproved of this practice due to the unfairness they found in NMET score interpretation and use. Luo and Xiao (2018) claimed that the varied test qualities of different NMET forms had diminished the NMET's potential to guard test takers' equal rights to higher education because they prompted favoritism and nepotism.

On the other hand, our study shows a contradictory finding. According to the surveyed students, the NEEA- and provincial self-developed NMET forms were both of high test qualities, and the accusation was utterly far-fetched that the self-developed forms were evidently easier than the NEEA-constructed ones. Although they did call for a nationwide unified NMET form out of their concern over the risk of regional discrimination, they cogently argued that this fairness issue was induced by the macro-level province-specific quota system for college enrollment, not by the test qualities of the micro-level NMET or any other *gaokao* subject. Every year, Chinese universities allocate a fixed admission quota for each provincial region, usually with a greater proportion of students originally from where the universities are based. Considering the uneven distribution of higher educational affordances across the Chinese mainland, this unequal quota admission policy is prone to intensify the competition for university seats among test takers from provincial areas with less educational resources or with fewer universities per capita. This is the very rationale behind our research participants' abovementioned second consideration for a nationally uniform NMET form.

However, it must be acknowledged that changing the college admissions quota system is not within the power of NMET or any other *gaokao* test developers. Given the current circumstances where each provincial region constitutes one single college enrollment unit, we maintain that the maximal extent of fairness the NMET can guarantee is that all *gaokao* candidates within each provincial area: (1) are granted equal status throughout the whole process of test design, administration, rating, and score use; and (2) have an opportunity to harness the cascading washback effects to enhance their English language development. For the former, Zhang's (2019) research reported test takers' satisfaction with the status quo, and for the latter issue, the present study found that students in general positively evaluated the effectiveness of the actual before-test NMET washback in promoting their senior high school English learning outcomes, though more work is needed to live up to students' expectations.

Looking to the future, we identify four directions for further research on NMET washback. First, while studying undergraduate students' senior high school experiences provides valuable insights, it is incomplete as it does not include those students not admitted to universities. Researching this cohort is daunting due to their dispersed nature and researchers' resource constraints. However, their inclusion is crucial as they constitute a substantial segment of senior high school students who may offer unique perspectives on NMET washback distinct from their university-bound peers.

Second, although students/test takers bear the most direct consequences of the NMET test results, their side of the story does not speak for other stakeholders. The perceptions and attitudes of other relevant groups (e.g., teachers, NMET developers, policymakers, and parents) should not be overshadowed in NMET washback research. Large-scale nationwide investigations are thus necessary to explore what these stakeholders think of the importance and status quo of NMET washback on senior high school students' English learning outcomes. This line of research may potentially contribute to a broader and clearer picture of the complex NMET mechanism in the Chinese mainland.

Moreover, although survey studies will remain essential, we encourage future researchers to mix methodologies (e.g., qualitative, naturalistic and ethnographic research, case studies) and to vary their methods of data collection (e.g., interviews, observations). In mixed-methods designs, the findings yielded from different methods can be triangulated and corroborated, and the strengths of both quantitative and qualitative methods can be built on while their weaknesses minimized. Therefore, washback researchers may consider incorporating mix-methods dimensions into their future studies for a richer and more nuanced understanding of NMET washback.

Finally, the variables influencing the NMET washback need to be inspected in a more situated and dynamic manner. For example, the present study found significant differences among students from disparate provincial areas regarding their perceptions of NMET washback on their listening and speaking learning outcomes in senior high school settings. Nevertheless, it is plausible that an underlying factor, whether the NMET listening and speaking subtests are mandatory, could underpin these observed. However, it is essential to recognize that these findings are based solely on inferential statistical analyses, while the true nature of these dynamics is likely multifaceted and warrants comprehensive inquiry into the contextualized and interconnected dynamics among all relevant variables to validate its significance. Additionally, factors in situ may interact with the changing circumstances while developing and varying through time. As a result, longitudinal data must be obtained to align with this perspective. The longitudinal approach readily lends itself to studying NMET washback in ways that manifest its complex and emergent nature and situate the realities firmly in context. With the content reform of the *gaokao* well underway, researchers' close ongoing engagement with the NMET washback over time will help students, teachers, and policymakers identify how the washback develops through interaction with the environment and how it may shift, so they can adjust learning, pedagogical, and reform strategies to best support students' English learning development for maximum language learning outcomes.

Abbreviations

ASL	Arabic as a second language
CET	College English Test
CET-SET	College English Test Spoken English Test
CSE	China's Standards of English Language Ability
EAP	English for Academic Purposes
EFL	English as a foreign language
GEPT	General English Proficiency Test
HKCEE	Hong Kong Certificate of Education Examination
IELTS	International English Language Testing System
MET	Matriculation English Test
MoE	Ministry of Education of People's Republic of China
NEEA	National Education Examinations Authority of China
NMET	National Matriculation English Test
OLS	Ordinary least squares
SBA	School-based assessment
TEM	Test for English Majors
TOEFL	Test of English as a Foreign Language
TOEFL iBT	TOEFL Internet-Based Test

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-024-00286-0.

Additional file 1: Appendix 1. Questionnaire (Excerpts, English version). **Appendix 2.** Parents' occupations and highest education levels. **Appendix 3.** Descriptives of survey respondents' responses to the 5-point Likert scale items.

Acknowledgements
The author would like to thank the anonymous reviewers for their comments on the earlier version of the manuscript.

Authors' contributions

The paper was single authored by the corresponding author.

Funding
The present study was supported by the project ***School Age Cohort Study of Brain and Mind Development in China (***STI 2030—Major Projects 2021ZD0200500); the project ***Regional Assessment of Education Quality (***23QYHX-A005); and the ***Fundamental Research Funds for the Central Universities (***2022NTSS32***).***

Availability of data and materials

Data are available upon reasonable request.

Declarations

Competing interests

The author declares no competing interests.

Received: 7 January 2024 Accepted: 26 March 2024
Published online: 15 April 2024

References

Alderson, J. C. (1986). Innovations in language testing. In M. Portal (Ed.), *Innovations in language testing* (pp. 93–105). NFER-NELSON.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279.

Chen, K., Wu, H., Li, X., & Qiao, H. (2019). Jiyu gaokao pingjia tixi de yingyuke kaoshi neirong gaige shishi lujing [The implementation path of the English examination content reform based on the Gaokao Assessment Framework]. *China Examinations, 12*, 33–37.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.

Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing, 28*(2), 221–249.

Cheng, L., & Curtis, A. (2010). The impact of English language assessment and the Chinese learner in China and beyond. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 267–273). Routledge.

Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly, 3*(1), 53–70.

Coleman, D., Starfield, S., & Hagan, A. (2003). The attitudes of IELTS stakeholders: Student and staff perceptions of IELTS in Australian, UK and Chinese tertiary institutions. *IELTS Research Reports, 5*, 159–235.

Dörnyei, Z. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). Routledge.

Ferman, I. (2004). The washback of an EFL National Oral Matriculation Test to teaching and learning. In L. Cheng & Y. Watanabe (with A. Curtis) (Eds.), *Washback in language testing: Research contexts and methods* (pp. 191–210). Lawrence Erlbaum Associates.

Green, A. (2003). *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university pre-sessional courses*. [Unpublished doctoral dissertation]. University of Surrey.

Green, A. (2006). Watching for washback: Observing the influence of the International English Language Testing System academic writing test in the classroom. *Language Assessment Quarterly, 3*(4), 333–368.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge University Press.

Gu, X. (2007). *Positive or negative—ositive or negativety Press.eparat*. Chongqing University Press.

Gu, X. (2014). *Daxue yingyu siliuji kaoshi fanbo xiaoying lishi yanjiu* [A longitudinal study of CET washback]. Sichuan University Press.

Gui, S., Li, X., & Li, W. (1988). Guangdongsheng yingyu biaozhunhua kaoshi shiyan de jiben jingyan [A reflection on experimenting with the standardized Guangdong version of the National Matriculation English Test]. *Curriculum, Teaching Material and Method, 11*, 10–17.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge University Press.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng & Y. Watanabe (with A. Curtis) (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97–111). Lawrence Erlbaum Associates.

Hughes, A. (1993). Backwash and TOEFL 2000. [Unpublished manuscript]. University of Reading.

Jin, Y. (2000). Daxue yingyu siliuji kaoshi kouyu kaoshi dui jiaoxue de fanbo zuoyong [The washback of CET-SET on English teaching]. *Foreign Language World, 4*, 56–61.

Li, X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development, 11*(5), 393–404.

Liu, Q. (2010). The National Education Examination Authority and its English language tests. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 29–43). Routledge.

Luo, J., & Xiao, Y. (2018). Gaofengxian yuyan ceshi de gongpingxing jianyan kuangjia yanjiu—Yi gaokao yingyu weili [Evaluating fairness of high-stakes language tests: An empirical approach based on the NMET]. *Foreign Language Research, 1*, 86–91.

Ma, J. (2019). Did test preparation practices for the college English test work? A study from Chinese students' perspective. In S. Papageorgiou & K. Bailey (Eds.), *Global perspective on language assessment* (pp. 169–182). Routledge.

Mackey, A., & Gass, S. (2022). *Second language research: Methodology and design* (3rd ed.). Routledge.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education* (pp. 83–121). University of Chicago Press.

Ministry of Education. (2018). *Putong gaozhong yingyu kecheng biaozhun (2017 nian ban)* [English curriculum standards for general senior high schools (2017 edition)]. People's Education Press.

Ministry of Education. (2022, October). *2023 nian putong gaodeng xuexiao zhaosheng quanguo tongyi kaoshi dagang* [Technical manual of 2023 University Entrance Examination to Higher Education]. Retrieved November 07, 2023, from http://www.moe.gov.cn/.

Ministry of Education., & State Language Affairs Commission. (2018). *Zhongguo yingyu nengli dengji liangbiao* [China's standards of English language ability]. Higher Education Press & Shanghai Foreign Language Education Press.

National Education Examinations Authority of China. (2019). *Zhongguo gaokao pingjia tixi* [The Gaokao Assessment Framework of the People's Republic of China]. People's Education Press.

Qi, L. (2004). *The intended washback effect of the National Matriculation English Test in China: Intentions and reality*. Foreign Language Teaching and Research Press.

Qi, L. (2011). Yuyan ceshi de fanbo xiaoying lilun yu shizheng yanjiu [A review of washback studies]. *Foreign Language Learning Theory and Practice, 4*, 23–28.

Shih, C. (2006). *Perceptions of the General English Proficiency Test and its washback: A case study at two Taiwan technological institutes* [Unpublished doctoral dissertation]. University of Toronto.

Shih, C. (2007). A new washback model of students' learning. *The Canadian Modern Language Review, 64*(1), 135–161.

Shih, C. (2009). How tests change teaching: A model for reference. *English Teaching: Practice and Critique, 8*(2), 188–206.

Shih, C. (2010). The washback of the General English Proficiency Test on university policies: A Taiwan case study. *Language Assessment Quarterly, 7*(3), 234–254.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning. NFLC Occasional Papers*. The National Foreign Language Center.

Shohamy, E. (2000). Using language tests for upgrading knowledge: The phenomenon, source and consequences. *Hong Kong Journal of Applied Linguistics, 5*(1), 1–18.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317.

State Council of China. (2014, September). *Guowuyuan guanyu shenhua kaoshi zhaosheng zhidu gaige de shishi yijian* [Suggestions of the State Council of China on deepening the reform of examination and enrollment system]. Retrieved from http://www.gov.cn.

Vongpumivitch, V. (2012). Motivating lifelong learning of English? Test takers' perceptions of the success of the General English Proficiency Test. *Language Assessment Quarterly, 9*(1), 26–59.

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study*. Educational Testing Service.

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change*. Educational Testing Service.

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 3, the role of the coursebook, phase 4, describing change*. Educational Testing Service.

Wang, H., & Zhan, X. (2011). Gaokao yingyu ceshi gongpingxing de duowei fenxi [Multidimensional analysis on fairness of English teaching in College Entrance Examination]. *Curriculum, Teaching Material and Method, 31*(5), 49–53.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318–333.

Watanabe, Y. (1997). *Washback effects of the Japanese university entrance examination: Classroom-based research*. [Unpublished doctoral dissertation]. Lancaster University.

Watanabe, Y. (2004a). Methodology in washback studies. In L. Cheng & Y. Watanabe with A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19–36). Lawrence Erlbaum Associates.

Watanabe, Y. (2004b). Teacher factors mediating washback. In L. Cheng & Y. Watanabe (with A. Curtis) (Eds.), *Washback in language testing: Research contexts and methods* (pp. 129–146). Lawrence Erlbaum Associates.

Xu, Q. (2012). Yingyu zhuanye baji kaoshi de fanbo zuoyong yanjiu—Dui waiyu zhuanjia he yingyu xueke fuzeren de yici diaocha [The washback effects of TEM 8—From the perspectives of foreign language experts and English discipline leaders]. *Foreign Language World, 3*, 21–31.

Zhang, H. (2019). *A washback study of the National Matriculation English Test in China: Test takers' perspective* [Unpublished doctoral dissertation]. Tsinghua University.

Zhang, H., & Bournot-Trites, M. (2021). The long-term washback effects of the National Matriculation English Test on college English learning in China: Tertiary student perspectives. *Studies in Educational Evaluation*, *68*, Article 100977.

Zhu, L., & Yang, A. (2004). *Zouhuorumo de yingyu* [Too much for English]. Hunan People's Publishing House.

Zou, S., & Xu, Q. (2014). Biaozhun canzhao kaoshi ji qi fanbo xiaoying—Yi TEM kaoshi weili [Criterion-referenced assessment and its washback effects: The case of TEM]. *Foreign Language Learning Theory and Practice, 1*, 42–48.

Zou, S., & Xu, Q. (2017). A washback study of the Test for English Majors for Grade Eight (TEM8) in China—From the perspective of university program administrators. *Language Assessment Quarterly, 14*(2), 140–159.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.