Language Testing in Asia

RESEARCH Open Access

A moderated mediation analysis of the relationship between a high-stakes English test and test takers' extracurricular English learning activities

Jing Zhang

Correspondence: ydhsxyr@gmail.com
Department of Linguistics, Sophia University, Yotsuya Campus, 7-1 Kioi-cho, Chiyoda-Ku, Tokyo 102-8554, Japan

Abstract

This study investigated the relationship between a large-scale and high-stakes English test and test takers' learning behavior. Specifically, it explored whether and how the National Matriculation English Test (NMET) influenced test takers' extracurricular English learning activities under the Chinese Mainland educational context. Based on Bandura's triadic reciprocal determinism theory, this study proposed a distal mediation model and employed covariance-based Structural Equation Modeling to test the model. The data were collected via a cross-sectional survey with 470 test takers. The results showed that test takers' perceptions of the examination exerted direct and indirect effects on their extracurricular English learning activities, and that test takers' perceived self-efficacy for self-regulated learning and academic achievement were two important factors mediating the relationship between their perceptions of the test and extracurricular learning. Furthermore, test takers' perceptions of the exam-approaching have diverse moderating effects on different mediation effects. This study suggests that introducing the triadic reciprocal determinism theory helps understand how an examination influences learning. It also highlights the role of test takers' perceptions of an examination and their perceived self-efficacy in predicting a test's impact on learning.

Keywords: The NMET, Test takers' extracurricular English learning activities, Self-efficacy for self-regulated learning, Self-efficacy for academic achievement, Covariance-based Structural Equation Modeling, Bootstrapping

Introduction

This study was conducted under the Chinese Mainland educational background, with a particular focus on Gaokao—the college entrance examination for the entire country. The competition of Gaokao is so fierce that the mass media usually compare the difficulty of taking Gaokao to "thousands of troops crossing one narrow bridge" (Shi & Jia, 2015). Additionally, the number of test takers has been increasing in recent years, which reached 10,710,000 in 2020, an increase of 400,000 over

last year (Ministry of Education of the People's Republic of China, 2020). Hence, Gaokao is undoubtedly a large-scale and high-stakes test for most test takers in the Chinese Mainland. The current study only focused on the English component of Gaokao—the National Matriculation English Test (NMET).

Despite the importance of the NMET, its impact on teaching and learning has not attracted enough attention (Dong, 2018; Zou & Dong, 2014). The NMET is designed to help universities select qualified students and to guide teaching and learning in senior high schools (Ministry of Education of the People's Republic of China, 2017). Thus, in this context, the impact of the NMET on teaching and learning deserves scrutiny. In the Chinese Mainland, since the inception of the impact studies of the NMET in 1990, its impact on teaching has been the predominant focus (e.g., Dong, 2014; Dong, 2018; Li, 1990; Qi, 2004). However, the impact of the NMET on learning has been under-investigated (Zou & Dong, 2014). Test takers are the most important stakeholders of a test (Green, 2013; Rea-Dickins, 1997) and test takers' perceptions of the test are of great importance because these exert influence on their learning behavior (Hughes, 1993). It is thus reasonable to infer that understanding the mechanism of test's impact on learning might help improve test takers' learning. Hence, this study aims to investigate the relationship between the NMET and test takers' learning, particularly their extracurricular English learning.

Literature review

In the field of language testing, a wealth of studies investigating test impact on learning have reported that test takers engaged in extracurricular English learning activities during test preparation (e.g., Sato, 2019; Zhan & Andrews, 2014). However, most studies merely focused on traditional test preparation behavior, such as doing past papers (e.g., Xie & Andrews, 2012), while only a few studies highlighted the importance of test takers' extracurricular English learning activities (TEELA) and the relationship between a large-scale and high-stakes examination and TEELA, and even fewer studies specifically addressed the issue of whether such a relationship changes with the exam time approaching.

TEELA is an important type of learning that deserves attention. It refers to the communicatively-oriented English learning activities test takers are engaged in outside the classroom, such as reading English novels or watching TED lectures. Compared with traditional learning activities that are typically assigned and supervised by teachers or schools, TEELA is usually autonomous and somewhat like amusements that might help students to relax from a mountain of schoolwork. TEELA is thus not a test preparation practice per se in a way that test takers work on past examination papers. Extracurricular learning activities are not only an important contributor to students' academic achievement (e.g., Cooper, Valentine, Nye, & Lindsay, 1999) but also a facilitating factor for improving their language skills (e.g., Cao, 2015; Huang & Naerssen, 1987; Marefat & Barbari, 2009; Pan, 2014). In the Chinese Mainland, it is also believed that extracurricular English learning activities are instrumental in helping students achieve their long-term learning goals and improving their comprehensive language skills (Cao, 2015; Liang, 2011). Moreover, NMET test designers also regard developing students' comprehensive language skills as their supreme goal (Ministry of Education of

the People's Republic of China, 2017). Hence, it is warranted to examine whether and how the NMET influences TEELA.

In terms of the relationship between a large-scale and high-stakes examination and TEELA, contradictory conclusions have been gained under various educational contexts. For example, Zhan and Andrews (2014) conducted a case study in the Chinese Mainland and concluded that undergraduate test takers engaged in TEELA at the early stage of test preparation, and they admitted that they did such activities due to the influence of the examination. On the contrary, Sato (2019) implemented an exploratory study in Japan and found that senior high school test takers engaged in TEELA due to their interest in English rather than test impact.

Studies investigating whether the relationship between the test and TEELA changes as the exam time approaches are rare. Most research employed univariate techniques such as *t* tests to examine whether the exam time approaching affects TEELA. For example, Pan (2014) reported that the frequency of college students' engaging in TEELA increased as the exam time approached. It appears that although researchers realized the exam time approaching might influence TEELA, its role in moderating the relationship between the test and TEELA has not aroused enough attention.

In the test impact literature, test takers' perceptions have typically been used as predictors to represent a test. For example, Xie and Andrews (2012) employed test takers' perceptions of test design and test use as the predicting variables to examine the relationship between the College English Test and test takers' test preparation behavior. The present study follows this practice—using test takers' perceptions of the NMET (TPN) as the predictor, which is defined as test takers' perceptions of the positive influence that the NMET exerts on their English learning. This definition is inspired by the idea that a well-designed test might motivate test takers to be engaged in learning activities that are beneficial to their long-term learning goals (Green, 2013). For students, a well-designed test might mean a test that exerts a good effect on learning. Cheng, Andrews, and Yu (2010) have used a similar construct to investigate test takers' perceptions of a newly-introduced test. Nevertheless, the construct was treated as an outcome variable in their research.

Another gap identified in impact studies regarding learning was that most research adopted qualitative methods (e.g., Sato, 2019; Zhan & Andrews, 2014), with a particular lack of confirmatory studies of mediating factors (Sato, 2019). The existing literature suggests that many mediating factors exist on the testing–learning path, and applying qualitative methods enables researchers to identify these factors (Watanabe, 2004; Xie, 2015). For example, Watanabe (2004) has summarized five types of mediating factors based on previous research, including test factors, prestige factors, personal factors, micro-context factors, and macro-context factors. However, these factors were underexplored (Xie, 2015), meaning that little has been investigated about their "relative importance" (Xie, 2015, p. 58), their relationships (Sato, 2019), and their generalizability to diverse situations. Thus, researchers are encouraged to employ "more sophisticated data collection and analysis methods" (Tsagari & Cheng, 2017, p. 368). Xie and Andrews (2012), for example, conducted a mediation analysis and showed that the expectation of success was a good mediator on the path from test taker perceptions of the examination to test preparation behavior. However, in their research, the construct of the expectation of success was measured by the self-efficacy scale, suggesting that

self-efficacy might be a good factor mediating the relationship between a test and test takers' learning. The mediating effect of self-efficacy accounting for the impact of test taker perceptions on their learning behavior is thus worth further scrutiny. Additionally, estimation methods of mediation effects employed in the existing impact research, such as the products of coefficients approach, were lack of statistical power (see Data analysis). Consequently, it is necessary to find a new approach to analyzing mediation effects.

Theoretical framework

This study introduced Bandura's triadic reciprocal determinism (TRD) theory (1986) to explain the process of the NMET's impact on learning.

TRD theory attempts to explain humans' learning behavior in the social environment. It proposes that environmental factors, personal factors, and behavior are independent of each other, but they interrelate with and determine each other (Bandura, 1986). Environmental factors refer to the external social events that greatly influence individuals, for example, the NMET is an influential environmental factor for test takers; personal factors, such as cognitive, emotional, and motivational factors, play a strong controlling and guiding role in human behavior (Guo & Jiang, 2008). The three elements do not always exert equivalent influence on each other, and their influences change due to different circumstances, individuals, and activities.

The TRD model involves three interactions: The interaction between the environment and the person describes that the environment interacts with human beliefs and cognitive competencies (Guo & Jiang, 2008). The interaction between the person and behavior refers to the interaction of human thoughts and actions. The interaction between the environment and behavior depicts that the environment influences human behavior, which in turn influences that environment. Thus, based on this model, the NMET, test takers' factors, and their learning behavior interrelate with each other. Specifically, there are interactions between the NMET and test takers' belief about the NMET, between test takers' thoughts and actions, and between test takers' learning behavior and certain aspects of the NMET. Besides, personal factors have been assumed to be mediating factors between a test and learning behavior (e.g., Watanabe, 2004); thus, it might be reasonable to infer that test takers' perceptions of the NMET exert an impact on the personal factors, and in turn influence their learning behavior.

Within the framework of the TRD theory, Bandura further explored the personal factors. Particularly, he highlights the importance of self-efficacy, a cognitive self-concept of the capabilities that "one can successfully execute the behavior required to produce desired outcomes" (Bandura, 1977, p. 193), because perceived self-efficacy is helpful in explaining a myriad of phenomena such as "changes in coping behavior produced by different modes of influence" (Bandura, 1982, p. 122). According to Bandura (1982), people first form their perceptions of the environment. Based on these perceptions, individuals appraise their efficacy. High self-percept of efficacy may encourage people to deploy their efforts to deal with the demands of the environment and in turn enhance their performance, while low self-percept of efficacy may lead people to maximize the potential difficulties, which in turn jeopardize their performance. Therefore, there is strong reason to suspect that under the context of testing, test takers may first have

their perceptions of the test, then evaluate their self-efficacy based on their perceptions, which may finally affect their learning behavior.

Self-efficacy is a multidimensional construct (Bandura, 1986), in which perceived self-efficacy for self-regulated learning (PSE-SRL) and academic achievement (PSE-AA) are two strong predictors for student academic learning and performance (Oliveira, Taveira, Porfeli, & Grace, 2018; Zimmerman, Bandura, & Martinez-Pons, 1992). PSE-SRL refers to the prediction of one's capabilities to actively and systematically use self-regulatory process to gain the desired learning outcome (Lee, Lee, & Bong, 2014). Self-regulated learners display "a high sense of efficacy in their capabilities, which influence their commitment to fulfilling these challenges" (Zimmerman et al., 1992, p. 664). PSE-AA is defined as the conviction that learners can successfully attain their desired academic achievement (Schunk, 1991). A high sense of PSE-AA motivates learners to deploy more efforts, persistence, and intrinsic interest in their learning and performance (Zimmerman et al., 1992). Additionally, PSE-SRL has been proved to predict PSE-AA (Lee et al., 2014; Zimmerman et al., 1992). However, the effects of these two kinds of self-efficacy in terms of improving students' extracurricular English learning and their mediating effects between testing and learning behavior were under-investigated within the field of language testing. Only Xie and Andrews (2012) have explored the mediating effect of self-efficacy, but the self-efficacy measure used in their research focused more on motivated learning strategy. Therefore, little attention has been devoted to the mediating role of PSE-SRL and PSE-AA. Thus, this study conducted a mediation analysis to explore the effects of these two types of self-efficacy and their relationship.

Conceptual model and research questions

Based on the TRD theory and related literature, this study proposes that TPN influences test takers' PSE-SRL and PSE-AA, which in turn affect their TEELA. This process is moderated by test takers' perceptions of exam-approaching. Specifically, the following conceptual model (Fig. 1) depicts the proposed theory:

Three research questions are included in this study:

1. Does TPN have a direct effect on TEELA?

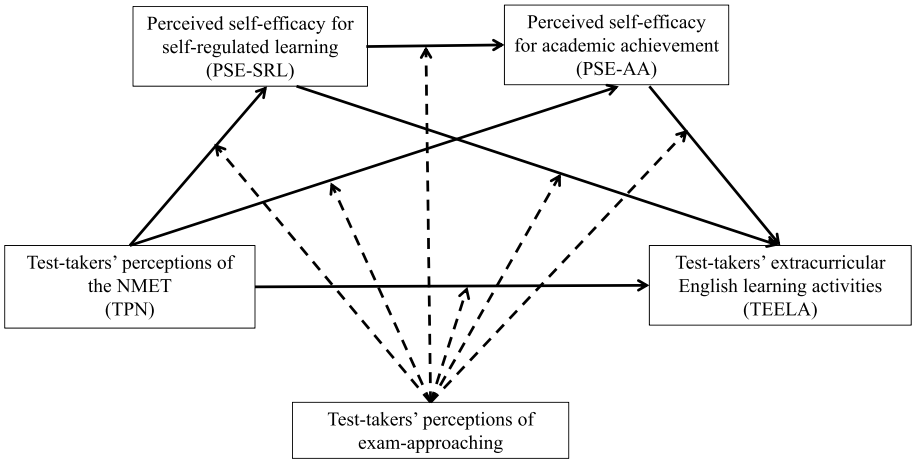


Fig. 1 Conceptual model

a. If this direct effect exists, will it change with the exam time approaching?

2. On the path from TPN to TEELA,
   a. Does PSE-SRL mediate the relationship between TPN and TEELA?
   b. Does PSE-AA mediate the relationship between TPN and TEELA?
   c. Does the TPN→PSE-SRL→PSE-AA→TEELA path exist?
3. Will test takers' perceptions of exam-approaching
   a. Moderate the indirect effect of TPN on TEELA through PSE-SRL?
   b. Moderate the indirect effect of TPN on TEELA through PSE-AA?
   c. Moderate the indirect effect of TPN on TEELA through PSE-SRL and PSE-AA?

Method

Research context and participants

The NMET aims to examine test takers' language knowledge and use (Ministry of Education of the People's Republic of China, 2019). In terms of language knowledge, test takers are required to master and use English phonetics, vocabulary, grammar, function-notion, and topics that they have learned. In terms of language use, the NMET examines test takers' ability from four perspectives: listening, reading, writing, and speaking. Table 1 describes the components of the NMET written test paper used in the province where the present study was conducted. All test takers are required to take the written test. On the contrary, the NMET spoken test is separate and optional. Typically, two types of students take this test: students wishing to apply for special majors such as foreign affairs and international law and students wishing to know their spoken English level. Test formats include reading a short passage aloud and answering the examiner's questions.

This research was conducted in an Eastern province in the Chinese Mainland. From five ordinary senior high schools (Table 2) in the capital city of the province, 470

Table 1 Components of the NMET written test

Part	Test content	Test format	Weighting	Percentage
I Listening comprehension	Section 1: Five short dialogues (five items)	Three-option MCQ	7.5	20%
	Section 2: Five long dialogues or monologues (15 items)	Three-option MCQ	22.5	
II Reading comprehension	Section 1: Four passages of varying length (15 items)	Four-option MCQ	30	27%
	Section 2: One passage with five missing sentences (five items)	Selecting-five-from-seven-option MCQ	10	
III Language knowledge use	Section 1: A cloze (20 items)	Four-option MCQ	30	30%
	Section 2: A passage with ten blanks, some of the blanks are followed by a cue word (10 items)	Blank filling	15	
IV Writing	Section 1: Proofreading a short text, with each line containing one error (10 items)	Error correction	10	23%
	Section 2: Writing a short passage of approximately 100 words based on the hypothetical situation provided in the instruction	Guided writing task	25	

Note: MCQ, multiple-choice questions

students were randomly selected for this study. Based on Hair Jr., Black, Babin, and Anderson's (2019) suggestion, a sample size of 470 is large enough for this study. There is no wide disparity among these schools in terms of teaching quality, school facilities, the minimum score of high school admission, and philosophies of schooling. All five English teachers agreed to include several randomly selected classes in the present study. Besides, random selection within the classes was performed by the author. Table 3 shows the demographic characteristics of these participants.

Instrumentation

A questionnaire (see Appendix), including four multi-item measures (31 items), was employed to assess the latent constructs in the conceptual model. All measures were revised from other researchers' scales so that they were originally developed in English. Having been examined and discussed by three experts, all the items were translated into Chinese via the translation–back translation procedure (Brislin, 1970). Before the formal data collection, at the end of 2019, a pilot study of 89 senior high school students from one middle school in the same province with the formal survey, was conducted to evaluate the quality of the research design and questionnaire items. No problematic items were identified based on the results of the item analysis.

Test takers' perceptions of the NMET

TPN was assessed by a nine-item scale adapted based on the "students' perception subscale" developed by Cheng et al. (2010) and the NMET syllabus issued in 2019. High TPN score means that test takers believe the NMET can influence their English learning positively. The respondents were asked to choose from a seven-point Likert scale ranging from 1, "strongly disagree", to 7, "strongly agree". The Cronbach's alpha (in the actual administration) for the TPN subscale was .928.

Test takers' perceived self-efficacy

Two subscales from the Multidimensional Scales of Perceived Self-Efficacy (Bandura, 1989, as cited in Williams & Coombs, 1996) were selected and revised for use in the present study: PSE-SRL and PSE-AA. The PSE-SRL subscale was composed of 10 items, measuring test takers' perceived ability to use diverse self-regulated learning strategies. The PSE-AA subscale consisted of six items assessing test takers' perceived capability to gain success in six aspects: English vocabulary, grammar, reading, listening, speaking, and writing. Participants rated the strength of their belief on a 7-point scale ranging from 1, "not well at all", to 7, "very well". The Cronbach's alphas of the PSE-SRL and PSE-AA subscales were .952 and .958, respectively.

Table 2 Description of the high schools

School	Number of the participants	Location of the school	Type of the school
1	46	Suburban	Public
2	110	Urban	Private
3	50	Urban	Public
4	180	Urban	Public
5	84	Urban	Public

Table 3 Demographic characteristics of the participants

Characteristics	Number	Percentage (%)
Grade		
1	255	54.3
2	115	24.5
3	100	21.3
Gender		
Female	297	63.2
Male	173	36.8
Living status		
Living in the dormitory	428	91.1
Living at home	42	8.9

Note: N = 470

Test takers' extracurricular English learning activities

TEELA was measured by a six-item subscale modified from the "test-related English activities outside school" subscale in the study of Cheng et al. (2010). Items in the TEELA scale measured test takers' frequency of engaging in TEELA in the past year. The items were responded to on a 7-point Likert scale with values varying from 1, "never", to 7, "every time". The Cronbach's alpha of this subscale was .940.

Test-takers' perceptions of exam-approaching

This construct is represented by three grades in high school. The higher the grade, the stronger test takers' perceptions or senses of the exam-approaching. Because the NMET is held at the end of senior three, the grade 3 students are the closest to the examination. As a consequence, compared with grade 1 and 2 students, grade 3 students face more pressure of Gaokao and spend more time and energy in test preparation (Cao, 2016), It is thus reasonable to infer that with the advance of grade, students' perceptions of the time of testing become increasingly intense.

Data collection

This study involved a cross-sectional survey conducted in the spring of 2020. To guarantee the reliability of the responses and absolute confidentiality, the participants were assured of anonymity, and they were ensured that only the researcher would see their responses. The survey was created and implemented with a widely used tool—WENJUANXING (http://www.wenjuanxing.com). One advantage of using WENJUANXING is that no missing data will be generated due to its prior setting (if respondents forget to fill in one item, they will be reminded to complete it; otherwise, they cannot continue with the questionnaire). Students who completed and successfully submitted the questionnaire joined in an online lucky draw immediately after their submission, and several types of awards were provided as a token of gratitude from the author.

Data analysis

Analytic strategy

This study employed the covariance-based Structural Equation Modeling (CB-SEM) technique to answer the research questions with Amos 24. CB-SEM is typically used to test process models developed by a theory (Hayes, 2009; Lei & Wu, 2007). When using CB-SEM, investigators do not find a model to fit the data (Kline, 2016), but test a theory via specifying a model depicting the relationships between the constructs that are described in that theory, with the constructs measured by valid observed variables (Hair Jr. et al., 2019). In doing so, researchers can "evaluate the validity of substantive theories with empirical data" (Lei & Wu, 2007, p. 33), which in turn helps develop a theory (Anderson & Gerbing, 1988). Hence, the present study employed CB-SEM to reveal what happened in the process of the NMET exerting influence on learning.

The maximum likelihood estimation method was employed because it has been known to gain more robust parameter estimates compared with other estimators (e.g., generalized least squares) (Curran, West, & Finch, 1996), even when the observed variables were not on a multivariate normal distribution (Iacobucci, 2010).

To answer the research questions, this study administered three analyses. Firstly, confirmatory factor analysis (CFA) was performed to assess the measurement model. Secondly, mediation analysis was conducted employing bootstrapping (Hayes, 2009) to answer research questions 1 and 2. Finally, the subgroup method and bootstrapping were applied to conduct a moderated mediation analysis to answer research question 3. All the bootstrapping procedures were conducted with 5000 bootstrap samples (Hayes, 2009).

Effect sizes were also discussed. Hedges' g was calculated to gauge how different groups of test takers varied (Ellis, 2010). Besides, Pearson product moment correlation coefficient (r) and coefficient of multiple determination (R^2) were applied to measure the strength of the relationships between constructs (Ellis, 2010).

Mediation analysis

A distal mediation model (Fletcher, 2006) was developed in this study, as illustrated in Fig. 2. *a1* and *a2* represent the path coefficients from *TPN* to *PSE-SRL* and *PSE-SRL* to *TEELA*, respectively. *b1* and *b2* represent the path coefficients from *TPN* to *PSE-AA* and *PSE-AA* to *TEELA*, respectively. *c* is the path coefficient from *PSE-SRL* to *PSE-AA*. *d* is the path coefficient from *TPN* to *TEELA*, representing the direct effect of *TPN* on

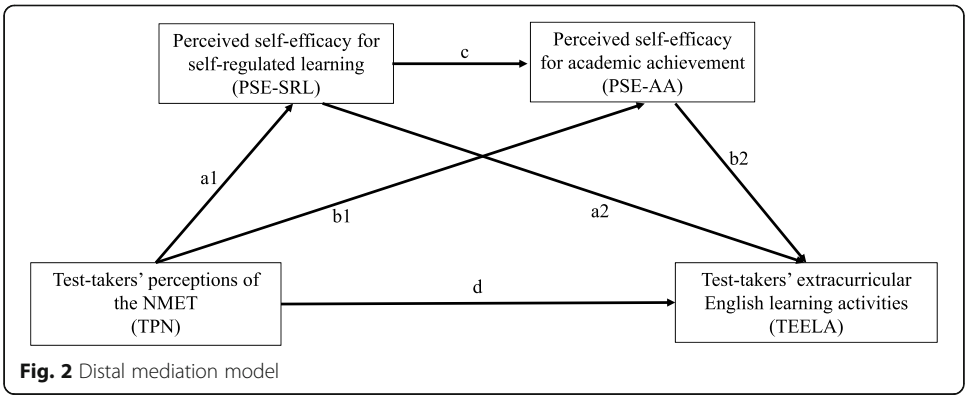


Fig. 2 Distal mediation model

TEELA. Three specific indirect effects (SIE) are included in this model: The product of *a1* and *a2* represents the mediation effect of *TPN* on *TEELA* through *PSE-SRL* (SIE 1). The product of *b1* and *b2* represents the indirect effect of *TPN* on *TEELA* via *PSE-AA* (SIE 2). The product of *a1*, *c*, and *b2* represents the distal mediation effect of *TPN* on *TEELA* through *PSE-SRL* and *PSE-AA* (SIE 3). The total indirect effect is quantified as SIE 1 + SIE 2 + SIE 3, while the total effect is quantified as SIE 1 + SIE 2 + SIE 3 + *d*.

The assessment of such a process model is mediation analysis, which allows researchers to understand by what means a predicting variable exerts its influence on an outcome variable (Preacher, Rucker, & Hayes, 2007). The mediation effect or indirect effect deserves proper attention, otherwise, "the relationship between two variables of concern may not be fully considered" (Raykov & Marcoulides, 2006, p. 7).

Diverse methods can be used to gauge the magnitude of indirect effects. Baron and Kenny's (1986) causal steps approach has been the most widely used one (Hayes, 2009; MacKinnon, Lockwood, & Williams, 2004). However, it has been criticized for the lowest statistical power (Fritz & MacKinnon, 2007; Hayes, 2009), and it is only applicable to the simple mediation model (Preacher et al., 2007). As a consequence, investigators usually adopt the Sobel test as a "supplement" (Hayes, 2009, p. 6) to the causal steps approach. Nevertheless, both of the causal steps approach and Sobel test are based on the premise that the product of *a1* and *a2* (or *b1* and *b2*) is normally distributed, which is difficult to achieve (Bollen & Stine, 1990; Preacher et al., 2007; Stone & Sobel, 1990). Thus, the present study introduced a cutting-edge technique—bootstrapping (Bollen & Stine, 1990; Hair Jr. et al., 2019; Hayes, 2009; Preacher et al., 2007) to assess mediation effects, which does not require the assumption of normal distribution (Hayes, 2009; Preacher et al., 2007).

Two forms of bootstrapping were adopted in this study: naive bootstrapping (Yung & Bentler, 1996) and Bollen–Stine bootstrapping (Bollen & Stine, 1992). The former was used to conduct a mediation analysis (Hayes, 2009; Preacher et al., 2007), and the latter was applied to modify the enlarged χ^2 due to multivariate nonnormality (Enders, 2005).

Moderated mediation

When the effect of an independent variable on a dependent variable varies due to different levels of a third variable, this variable is called a moderator (Baron & Kenny, 1986; Edwards & Lambert, 2007; James & Brett, 1984). As mediation analysis has aroused considerable attention, many researchers show interest in the condition under which an indirect effect occurs, which is thus referred to as conditional indirect effects (Preacher et al., 2007) or moderated mediation (James & Brett, 1984).

The most widely used method to examine moderated mediation is to analyze the mediation effect separately at each level of the moderator (Fabrigar & Wegener, 2014), which is called the subgroup approach (Edwards & Lambert, 2007). Following Preacher et al.'s (2007) suggestion, within each subgroup (grades 1, 2, and 3), mediation effects were estimated with the bootstrapping procedure.

Results

Data examination

To ensure the quality of CFA, outliers and distributional assumptions were examined first (Jackson, Gillaspy Jr., & Purc-Stephenson, 2009). Seven cases were judged to be outliers based on Mahalanobis *d* square values (Byrne, 2016) and were deleted from further analysis. Then multivariate normality was examined, which is the prerequisite of the maximum likelihood estimation (Byrne, 2016; Curran et al., 1996). Although all the observed variables exhibited univariate normality, the critical ratio of multivariate kurtosis value was above 5.00 (c.r. = 99.291), indicating that the data were multivariate nonnormal (Bentler, 2005), which may mislead the researcher to reject the correct model (Curran et al., 1996; Lei & Wu, 2007). Byrne (2016) thus recommended that researchers "correct the test statistic, rather than use a different mode of estimation" (p. 124). Hence, Bollen–Stine bootstrapping was applied to re-estimate chi-square and standard error (Bollen & Stine, 1992; Enders, 2005; Lei & Wu, 2007), which might help "gain insight into the behavior of the test statistic with nonnormal data" (Bollen & Stine, 1992, p. 229).

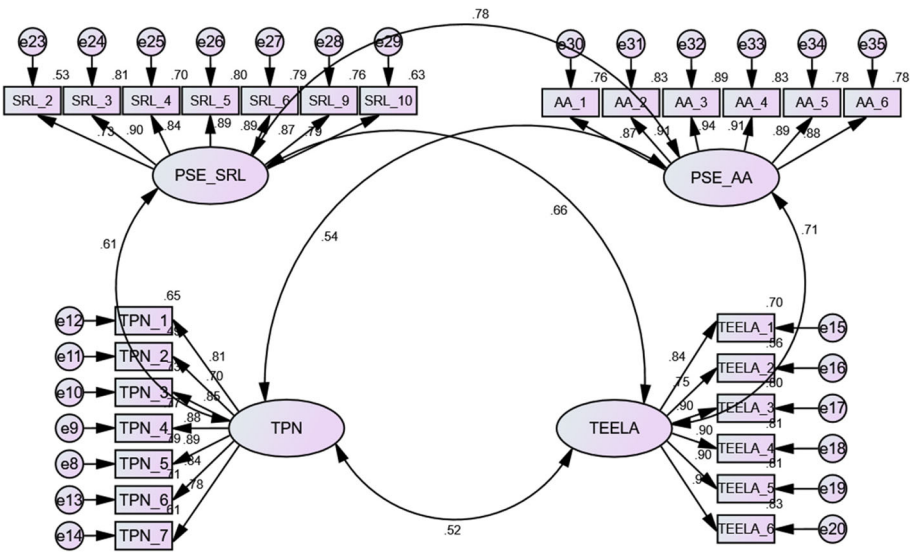
Measurement model

Before analyzing the structural model, the measurement model should be carefully tested to guarantee that all the observed variables reflect the desired latent constructs (Anderson & Gerbing, 1988; Jackson et al., 2009) and to determine how well the theoretically specified factor structures fit the sample data (Hair Jr. et al., 2019).

Following Hair Jr. et al.'s suggestion (2019), before formally assessing the measurement model, the diagnostic information from a preliminary CFA was used to modify the model slightly and to improve the quality of the model. Five problematic indicators (see Appendix) were identified. They exhibited the possibility of cross-loadings and error term correlations, which "would be inconsistent with the theoretical basis of CFA and SEM in general" (Hair Jr. et al., 2019, p. 678). After carefully considering the face validity and discussing with experts many times, the author decided to delete the five indicators from further analysis. The following section reported the results of assessing measurement model validity, including fit and construct validity.

Firstly, the fit validity was examined. Following Hair Jr. et al.'s (2019) and Jackson et al.'s (2009) suggestions, this study reported the following fit indices: chi-square value, relative chi-square (χ^2/df), root mean square error of approximation (RMSEA), Tucker Lewis Index (TLI), and comparative fit index (CFI). A relative chi-square of 3.0 or less is considered good, RMSEA values of lower than .08 are associated with good fitting, and TLI and CFI values that approach 1.0 are considered good (Hair Jr. et al., 2019). The model with 26 measured variables (Fig. 3) yielded a Bollen–Stine χ^2 of 424.274 with 293 degrees of freedom, a relative chi-square of 1.45, an RMSEA of .03, a TLI of .99, and a CFI of .99, which were highly suggestive that the specified factor structure fit the sample data reasonably well.

Then, the construct validity was evaluated (Table 4), which was the main target of CFA (Hair Jr. et al., 2019). All the standardized factor loadings were above .50 and significant ($p < .001$), meaning that the items were ideally convergent on their corresponding latent construct (Hair Jr. et al., 2019). Besides, all the AVE values were above .50, which was suggestive of adequate convergence (Hair Jr. et al., 2019). Further, all the SMC values were above .36, indicating that all the items were reliable (Fornell & Larcker, 1981). The



Note. *N* = *463*. *TPN* = test-takers' perceptions of the NMET; *PSE-SRL* = perceived self-efficacy or self-regulated learning; *PSE-AA* = perceived self-efficacy for academic achievement; *TEELA* = test-takers' extracurricular English learning activities. Coefficients presented are standardized coefficients.

Fig. 3 Measurement model

composite reliability of greater than .70 rendered enough evidence of good reliability, which suggested appropriate internal consistency within every construct (Hair Jr. et al., 2019). Table 5 contains the result of testing the discriminant validity. Following Hair Jr. et al.'s (2019) suggestion, the discriminant validity was assessed by comparing "the AVE values for any two constructs with the square of the correlation estimate between these two constructs" (p. 677). Thus, the square roots of AVEs were calculated and compared with correlation estimates. All square roots of AVEs were greater than the corresponding Pearson correlation coefficients, indicating that every construct was distinct from each other.

Overall, the results of the CFA showed that the specified measurement model fit well with the sample data, which provided a basic and vital premise for the subsequent structural model analysis (Hair Jr. et al., 2019).

Structural model

This section summarized the results of testing the proposed structural theory, which focused on examining the overall structural model fit and the hypothesized structural relationships between constructs. The structural model yielded a Bollen–Stine χ^2 of 424.274 with 293 degrees of freedom, a relative chi-square of 1.45, an RMSEA of .03, a TLI of .99, and a CFI of .99, indicating that the hypothesized structure adequately fit the observed covariance matrix.

Mediation analysis

Figure 4 illustrates the standardized path estimates and R^2 of the hypothesized model. All the path coefficients were statistically significant ($p < .05$), indicating

Table 4 The results of construct validity

Construct	Indicator	Test of significance				Factor loading (Std.)	Item reliability (SMC)	Composite reliability	Convergent validity (AVE)
		Unstd.	SE	*t* value	*p*				
TPN	TPN_1	1.000				.808	.653	.936	.677
	TPN_2	.894	.054	16.545	***	.699	.489		
	TPN_3	.937	.043	21.965	***	.852	.726		
	TPN_4	1.056	.047	22.595	***	.876	.767		
	TPN_5	1.048	.045	23.079	***	.888	.789		
	TPN_6	.948	.044	21.313	***	.842	.709		
	TPN_7	.947	.050	19.075	***	.778	.605		
TEELA	TEELA_1	1.000				.837	.701	.948	.751
	TEELA_2	.848	.044	19.092	***	.751	.564		
	TEELA_3	1.184	.047	25.409	***	.896	.803		
	TEELA_4	1.213	.048	25.522	***	.899	.808		
	TEELA_5	1.204	.047	25.502	***	.898	.806		
	TEELA_6	1.263	.048	26.089	***	.909	.826		
PSE-AA	AA_1	1.000				.870	.757	.963	.811
	AA_2	1.000	.035	28.896	***	.910	.828		
	AA_3	.994	.032	31.403	***	.943	.889		
	AA_4	1.007	.035	28.922	***	.911	.830		
	AA_5	.971	.036	27.230	***	.886	.785		
	AA_6	.980	.036	26.945	***	.881	.776		
PSE-SRL	SRL_2	1.000				.731	.534	.947	.718
	SRL _3	1.028	.052	19.934	***	.901	.812		
	SRL _4	.963	.052	18.368	***	.836	.699		
	SRL _5	1.084	.055	19.723	***	.892	.796		
	SRL _6	1.080	.055	19.691	***	.891	.794		
	SRL _9	1.041	.054	19.226	***	.871	.759		
	SRL _10	1.113	.064	17.383	***	.794	.630		

Note: *N* = 463. *TPN*, test takers' perceptions of the NMET; *TEELA*, test takers' extracurricular English learning activities; *PSE-AA*, perceived self-efficacy for academic achievement; *PSE-SRL*, perceived self-efficacy for self-regulated learning. *AVE*, average variance extracted; *SMC*, squared multiple correlations; *Std*., standardized; *Unstd*., unstandardized; *SE*, standard error.
***$p < .001$.

Table 5 Descriptive statistics, correlation matrix and discriminant validity

Construct	Mean	S.D.	TPN	PSE-SRL	PSE-AA	TEELA
TPN	5.862	1.296	**(.823)**			
PSE-SRL	5.172	1.239	.606	**(.847)**		
PSE-AA	4.636	1.472	.538	.778	**(.901)**	
TEELA	4. 197	1.820	.521	.656	.706	**(.867)**

Note: N = 463. On the diagonal are the square roots of AVEs that are shown in parentheses in boldface. Below the diagonal are the Pearson correlation coefficients between the latent constructs. *TPN*, test takers' perceptions of the NMET; *PSE-SRL*, perceived self-efficacy for self-regulated learning; *PSE-AA*, perceived self-efficacy for academic achievement; *TEELA*, test takers' extracurricular English learning activities

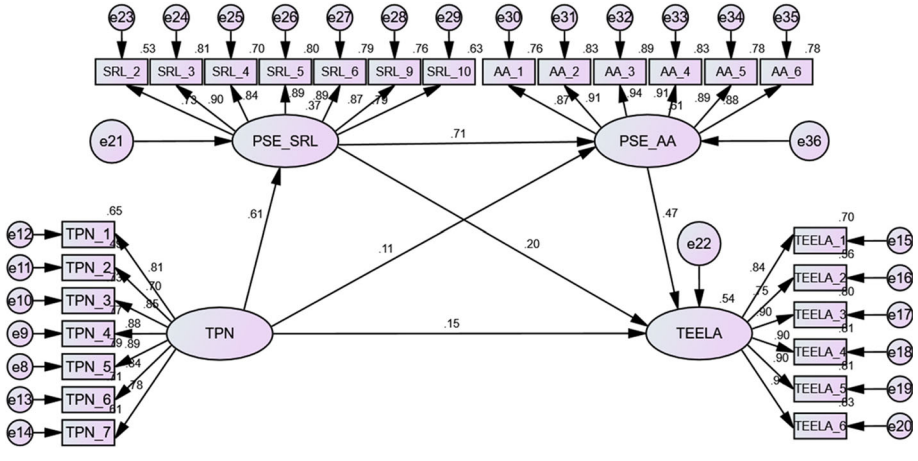
that all hypothesized relationships between constructs were supported. The R^2 for TEELA was .54, suggesting that the structural model explained 54% of the variance in TEELA. Table 6 summarizes the results of the mediation analysis. Five thousand bootstrapping with 95% confidence revealed that the direct path from TPN to TEELA was statistically significant (B = .179; p < .01). Additionally, TPN had an indirect, statistically significant, positive effect on TEELA via PSE-SRL (SIE 1) (B = .151; p < .01) or PSE-AA (SIE 2) (B = .062; p < .05). Besides, TPN also had an indirect, statistically significant, positive relationship with TEELA via PSE-SRL and PSE-AA (SIE 3) (B = .252; p < .001). All of the bootstrapping confidence interval ranges did not include zero, thus further proving that TPN had direct and indirect effects on TEELA, which also indicated the hypothesized model was a partial mediation model (Hair Jr. et al., 2019).

Finally, all possible pairwise comparisons among the three SIEs were examined to explore their relative importance, showing that only SIE 2 and SIE 3 was significantly different (SIE_{diff} = − .191; p = .000), while there was no statistically significant difference between SIE 1 and 3 (SIE_{diff} = − .102; p = .215), SIE 1 and 2 (SIE_{diff} = − .089; p = .219).

Moderated mediation analysis

As shown in Table 7, the moderated mediation analysis revealed that the total indirect effect and SIE 3 were statistically significant within each grade. However, neither SIE 1 nor SIE 2 was significant except for SIE 1 in grade 3 (B = .335; p < .001). The SIE comparison within each grade showed that there was no significant difference between SIE 1 and SIE 2 in three grades. SIE 1 and SIE 3 differed significantly (SIE_{diff} = − .345 and .205, respectively; p < .05) in grades 1 and 3. SIE 2 and SIE 3 differed significantly (SIE_{diff} = .364; p < .001) in grade 1.

Table 8 summarizes the results of the comparison of the indirect and direct effects among three grades. Despite no significant difference existing among the three grades



Note. N = *463*. *TPN* = test-takers' perceptions of the NMET; *PSE-SRL* = perceived self-efficacy for self-regulated learning; *PSE-AA* = perceived self-efficacy for academic achievement; *TEELA* = test-takers' extracurricular English learning activities.

Fig. 4 Structural Equation Modeling of the Hypothesized Model with Standardized Coefficients and R^2

Table 6 Direct, indirect, and total effects of the hypothesized model

Path	Point estimate (Std.)	Point estimate (Unstd.)	SE	*Z*	Bootstrapping with 95% confidence				Two-tailed significance
					Bias-corrected		Percentile		
					Lower	Upper	Lower	Upper	
Specific indirect effect									
SIE 1	.122	.151	.058	2.603	.039	.268	.041	.269	**
SIE 2	.050	.062	.031	2.000	.006	.130	.005	.127	*
SIE 3	.204	.252	.048	5.250	.169	.358	.169	.356	***
Total indirect effect									
TPN→TEELA	.376	.464	.062	7.484	.353	.592	.355	.596	***
Direct effect									
TPN→TEELA	.145	.179	.064	2.797	.057	.311	.051	.303	**
Total effect									
TPN→TEELA	.521	.644	.060	10.733	.537	.769	.533	.765	***

Note. 5000 bootstrap samples. $N = 463$. *Std.*, standardized; *Unstd.*, unstandardized; *TPN*, test takers' perceptions of the NMET; *TEELA*, test-takers' extracurricular English learning activities; *SIE 1*, TPN→PSE-SRL→TEELA; *SIE 2*, TPN→PSE-AA→TEELA; *SIE 3*, TPN→PSE-SRL→PSE-AA→TEELA
* $p < .05$, ** $p < .01$, *** $p < .001$

in terms of the direct effect and total indirect effect, grades 1 and 3 differed significantly in terms of SIE 1 and SIE 3 ($SIE_{diff} = -.307$ and .242, respectively; $p < .05$). The effect sizes were medium for the difference in SIE 1 (Hedges' $g = .224$) and small for that in SIE 3 (Hedges' $g = .129$).

Discussion

Research question 1 asks: "Does TPN have a direct effect on TEELA?" This study shows that TPN has a direct and positive effect on TEELA, suggesting that test takers who believe that the more positive impact the NMET has on their English learning, the more frequently they participate in extracurricular English learning activities. This is consistent with Cheng et al.'s (2010) finding that students who believed that the test had positive effects on their learning tended to engage in extracurricular English learning activities more frequently than those who held the opposite belief. Besides, this finding also partially coincides with Zhan and Andrews' (2014) conclusion that the College English Test drove test takers to engage in out-of-class English learning activities. Based on Cohen's (1988) benchmark, TPN is closely related to TEELA ($r = .521$, large effect size), but the path coefficient from TPN to TEELA is small ($\beta = .145$; $p < .01$), indicating that

Table 7 Moderated mediation analysis: point estimates and SIE comparison within each grade

Grade	Point estimate of the indirect effect				SIE comparison		
	Total indirect effect	SIE 1	SIE 2	SIE 3	SIE1–SIE2	SIE1–SIE3	SIE2–SIE3
1	.409***	.028 (ns)	.009 (ns)	.373***	.019 (ns)	− .345*	− .364***
2	.428***	.112 (ns)	.090 (ns)	.226***	.022 (ns)	− .114 (ns)	− .136 (ns)
3	.563***	.335***	.098 (ns)	.130***	.237 (ns)	.205*	− .032 (ns)

Note: 5000 bootstrap samples. $N = 463$. *SIE 1*, TPN→PSE-SRL→TEELA; *SIE 2*, TPN→PSE-AA→TEELA; *SIE 3*, TPN→PSE-SRL→PSE-AA→TEELA
* $p < .0$, *** $p < .001$

Table 8 Moderated mediation analysis: indirect and direct effect comparison among the three grades

Grade	Estimate	Bootstrapping with 95% confidence					
		Bias-corrected			Percentile		
		Lower	Upper	*p* value	Lower	Upper	*p* value
Total indirect effect comparison							
G1–G2	− .019	− .355	.256	.861	− .341	.264	.926
G1–G3	− .154	− .580	.118	.262	− .536	.145	.350
G3–G2	.135	− .210	.579	.412	− .234	.539	.473
SIE 1 comparison							
G1–G2	− .084	− .399	.195	.583	− .414	.183	.529
G1–G3	**− .307***	− .706	− .046	**.022**	− .658	.017	**.036**
G3–G2	.223	− .052	.622	.100	− .107	.540	.174
SIE 2 comparison							
G1–G2	− .081	− .271	.056	.250	− .268	.059	.272
G1–G3	− .090	− .286	.042	.194	− .268	.054	.248
G3–G2	.008	− .191	.212	.909	− .199	.209	.951
SIE 3 comparison							
G1–G2	.147	− .112	.379	.268	− .077	.407	.182
G1–G3	**.242***	.018	.482	**.038**	.003	.471	**.047**
G3–G2	− .096	− .341	.119	.304	− .303	.156	.462
Direct effect comparison							
G1–G2	− .005	− .302	.325	.965	− .303	.321	.976
G1–G3	− .005	− .291	.313	.959	− .297	.307	.985
G3–G2	.000	− .327	.324	.996	− .326	.325	.998

Note: 5000 bootstrap samples. $N = 463$. SIE 1, TPN→PSE-SRL→TEELA; SIE 2, TPN→PSE-AA→TEELA; SIE 3, TPN→PSE-SRL→PSE-AA→TEELA

*$p < .05$

there exist mediating factors between the two constructs, which also suggests that educators should attach great importance to test takers' perceptions of a test due to its potential in predicting and facilitating their extracurricular learning behavior. Specifically, test designers should communicate with test takers effectively and regularly. In doing so, they can understand test takers' ideas and accordingly provide helpful suggestions with students to guide their extracurricular learning, which may ultimately facilitate their academic achievement and language skills.

Research question 1 also asks: "If this direct effect exists, will it change with the exam time approaching?" Results show that the direct effect of TPN on TEELA does not change as the exam time approaches (Table 8). On the other hand, the indirect effect of TPN on TEELA via PSE-SRL and PSE-AA (SIE 3) decreases as the exam time is imminent (Table 7). This finding is consistent with Zhan and Andrews's (2014) conclusion that the frequency of college students participating in TEELA dropped as the exam time approached. However interestingly, the indirect effect of TPN on TEELA via PSE-SRL (SIE 1) increases with the exam time approaching. These findings indicate that test takers' perceptions

of exam-approaching plays a complex moderating role in the relationship between the test and extracurricular learning. Specifically, the exam time approaching exerts different influences on the direct and indirect effects of TPN on TEELA. Further investigations are thus needed to explore the moderating role of the exam time approaching.

Research question 2 is about how TPN exerts influence on TEELA. In this study, all the three mediation effects are statistically significant, indicating that PSE-SRL and PSE-AA might be useful and important mediators to explain how TPN affects TEELA, which is helpful in understanding the mechanisms of the test impact process. However, the standardized effect size of TPN→PSE-AA→TEELA (SIE 2) path is very small (β = .050; p < .05), and this path is not significant in three grades (Table 7), indicating that PSE-AA might not serve as an independent mediator to account for the TPN–TEELA relationship.

The SIE comparison shows that there is a significant difference between SIE 2 and SIE 3, suggesting that the SIE 3 path might be more important than the SIE 2 path when explaining how TPN affects TEELA. Specifically, test takers believing an examination influences their learning positively tend to have a high sense of self-regulated learning efficacy, driving them to take diverse self-regulated learning strategies, which in turn motivates them to be more confident about their capabilities to gain academic success and finally engage in out-of-class English learning activities frequently. On the SIE 3 path, the PSE-SRL is predictive of PSE-AA (β = .713; p < .001) and the effect size of the strength of their relationship is large (r = .778). Namely, learners with higher PSE-SRL tend to have higher PSE-AA, which suggests that educators should pay great attention to the importance of student PSE-SRL. This finding is consistent with Zimmerman et al.'s (1992) conclusion that PSE-SRL was predictive of PSE-AA (β = .512; p < .05).

The specified model explains 54% of the variance in TEELA, representing a large effect size, which shows that the selected factors make a significant contribution to TEEL A. Besides, the hypothesized model is a partial mediation one, indicating that there might be other mediators on the path from TPN to TEELA, which coincides with Xie and Andrews's (2012) conclusion that there were other mediating factors on the path from testing to learning. Further research is thus needed to explore other mediators (e.g., learner interest or test takers' anxiety) explaining how an examination affects extracurricular learning.

Research question 3 is concerned with the moderating effect of test takers' perceptions of exam-approaching. According to the moderated mediation analysis, although there is no significant difference in SIE 2 among three grades, grades 1 and 3 exhibit significant differences in the SIE 1 and SIE 3 (Table 8), suggesting that with the advance of grade, SIE 1 and SIE 3 change (Table 7), in which SIE 1 increases moderately (Hedges' g = .224, medium effect size) and SIE 3 decreases slightly (Hedges' g = .129, small effect size). Specifically, as exam time approaches, test takers who believe the NMET exerts a positive impact on their English learning are more confident about their ability to self-regulate learning strategically, which in turn motivates them to engage in extracurricular English learning activities more frequently. This is partially consistent

with Zimmerman and Martinez-Pons' (1990) finding that learners with higher PSE-SRL used learning strategies much greater than those with lower PSE-SRL. Additionally, the TPN→PSE-SRL→TEELA path is significant only in grade 3, suggesting that learners gradually become self-regulated with the exam approaching, which in turn motivates them to adopt diverse learning strategies. Further investigations, particularly longitudinal studies, are thus recommended to explore the mediating role of the PSE-SRL on the path from a test to test takers' learning behavior.

On the other hand, the TPN→PSE-SRL→PSE-AA→TEELA path is always statistically significant across the three grades, suggesting that this path might be the most effective one when explaining how TPN influences TEELA. As mentioned earlier, the effect of this path decreases with exam time approaching, which may be because students invest more and more energy in traditional test preparation activities as the exam time is imminent. More studies are still needed to explore why the strength of the relationship between TPN and TEELA via PSE-SRL and PSE-AA became weaker as the exam time approached.

Conclusion

This study was the initial effort to address the issue of whether, how, and when the NMET affects TEELA. The proposed model fit the obtained data reasonably well and explained a large proportion of the variance in TEELA, indicating that introducing the TRD theory provides enlightenment for understanding the mechanism of test's impact on learning. Additionally, this study provides empirical evidence for the hypothesis that many mediating factors might exist on the testing–learning path. The mediation effects of these mediators might diversify with the exam time approaching, which confirms that the mechanism of test impact is a highly complex process (Tsagari & Cheng, 2017) that calls for further investigation.

There were several limitations in this study. Firstly, this was a cross-sectional research under the educational context of the Chinese Mainland, and all the participants were from ordinary high schools. Thus, it should be cautious when generalizing the results to different educational settings. Secondly, this study gauged the effect sizes via Cohen's (1988) benchmarks, which should be the last choice when discussing effect sizes (Ellis, 2010). Durlak (2009) once pointed that rather than applying Cohen's benchmarking effect sizes as iron-clad criteria, researchers should examine the effect sizes obtained in prior relevant studies. However, in the test's impact literature, there is not enough previous related research to refer to when discussing effect sizes. Conducting more quantitative studies concerning test's impact on learning is thus warranted to help other investigators better understand the practical importance of the factors of concern. Finally, all data were from a self-reported questionnaire. It might be better to triangulate the findings with various techniques, which may further enrich the findings.

Appendix

Table 9 Student questionnaire

Dimension	Item
TPN	1. I think the NMET encourages me to read English books.
	2. I think the NMET encourages me to watch English movies or TV programs.
	3. I think the NMET encourages me to memorize vocabularies.
	4. I think the NMET encourages me to practice English writing.
	5. I think the NMET motivates me to learn English hard.
	6. I think the NMET helps me to become an independent learner.
	7. I think the NMET helps to improve my speaking skills.
	8. I think the NMET helps to improve my listening skills. *
	9. I think the NMET helps to improve my reading skills. *
PSE-AA	10. How well can you learn English grammar?
	11. How well can you learn vocabulary?
	12. How well can you learn reading skills?
	13. How well can you learn writing skills?
	14. How well can you learn listening skills?
	15. How well can you learn speaking skills?
PSE-SRL	16. How well can you finish homework assignments by deadlines? *
	17. How well can you study when there are other interesting things to do? (e.g., playing games on mobile phones)
	18. How well can you concentrate on school subjects?
	19. How well can you take notes of class instruction?
	20. How well can you plan your school work?
	21. How well can you organize your school work?
	22. How well can you remember information presented in class and textbooks? *
	23. How well can you arrange a place when you are out of school to study without distractions? *
	24. How well can you motivate yourself to do school work?
	25. How well can you participate in English class discussions?
TEELA	In the past 1 year, …
	26. I read English materials (e.g., books/magazines/ newspapers) out of school.
	27. I spend time on English entertainment outside school (for example, listened to English songs or watch English programs).
	28. I browse English websites out of school.
	29. I write in English in my free time, for example, writing blogs/emails/letters/ diaries in English.
	30. I practice to talk to others (e.g., my parents and friends) in English.
	31. I take part in out-of-class English events (e.g., English corner, English debate).

Note: * Represents deleted items.

Abbreviations

AVE: Average variance extracted; CB-SEM: Covariance-based structural equation modeling; CFA: Confirmatory factor analysis; CFI: Comparative fit index; NMET: National Matriculation English Test; PSE-AA: Perceived self-efficacy for academic achievement; PSE-SRL: Perceived self-efficacy for self-regulated learning; RMSEA: Root mean square error of approximation; SE: Standard error; SIE: Specific indirect effect; SMC: Squared multiple correlations; Std.: Standardized; TEELA: Test takers' extracurricular English learning activities; TLI: Tucker Lewis Index; TPN: Test takers' perceptions of the NMET; TRD theory: Triadic reciprocal determinism theory (Bandura, 1986); Unstd.: Unstandardized

Acknowledgements

I would like to thank my supervisor Professor Yoshinori Watanabe, my peer PhD students Ms. Makiko Kato and Ms. Makiko Habu. Also, I would like to thank my families. Finally, I would like to thank all the respondents who filled in the questionnaire.

Author's contributions

Jing Zhang performed the research and wrote this manuscript independently. The author(s) read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The author declares that she has no competing interests.

Received: 2 November 2020 Accepted: 28 February 2021
Published online: 08 April 2021

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–422.
Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215.
Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122–147.
Bandura, A. (1986). *Social foundations of thought and action: a social cognition theory*. Englewood Cliffs, NJ: Prentice-Hall.
Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.
Bentler, P. M. (2005). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software.
Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140.
Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205–229. https://doi.org/10.1177/0049124192021002004.
Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross Cultural Psychology*, *1*(3), 185–216.
Byrne, B. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*, (3rd ed.,). New York, NY: Routledge.
Cao, D. (2016). A reflection on senior high school student extracurricular English learning activities. *New Education Era Electronic Journal*, *22*, 60–60.
Cao, W. (2015). A preliminary discussion concerning senior high school student extracurricular English learning activities. *Middle School Curriculum Guidance*, *9*, 117–118.
Cheng, L., Andrews, S., & Yu, Y. (2010). Impact and consequences of school-based assessment (SBA): Students' and parents views of SBA in Hong Kong. *Language Testing*, *28*(2), 221–249.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed.,). Hillsdale, NJ: Lawrence Erlbaum.
Cooper, H., Valentine, J. C., Nye, B., & Lindsay, J. J. (1999). Relationships between five after-school activities and academic achievement. *Journal of Educational Psychology*, *91*(2), 369–378.
Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29.
Dong, L. (2014). A study of the washback effect of the NMET in Beijing on English language teaching and learning in the senior middle school (Doctoral dissertation). Retrieved from CNKI.
Dong, M. (2018). NMET washback on high school English classroom teaching. *Basic Foreign Language Education*, *20*, 25–32.
Durlak, J. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928.
Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*, 1–22.
Ellis, P. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.
Enders, C. K. (2005). An SAS Macro for implementing the modified Bollen-Stine bootstrap for missing data: Implementing the bootstrap using existing structural equation modeling software. *Structural Equation Modeling*, *12*(4), 620–641.
Fabrigar, L. R., & Wegener, D. T. (2014). Exploring causal and noncausal hypotheses in nonexperimental data. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology*, (2nd ed., pp. 936–990). Cambridge: Cambridge University Press.

Fletcher, T. (2006). Methods and approaches to assessing distal mediation [Paper presentation]. In *66th annual meeting of the Academy of Management*. Atlanta, GA: United States.
Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312.
Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239.
Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, *13*(2), 39–51.
Guo, B., & Jiang, F. (2008). *Self-efficacy theory and it's application*. Shanghai: Shanghai Educational Publishing House.
Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*, (8th ed.,). UK: Cengage Learning.
Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the New Millennium. *Communication Monographs*, *76*(4), 408–420.
Huang, X. H., & Naerssen, M. V. (1987). Learning strategies for oral communication. *Applied Linguistics*, *8*(3), 287–307.
Hughes, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript. Reading, U.K.: University of Reading.
Iacobucci, D. (2010). Structural equations modeling: Fit indices, ample size, and advanced topics. *Journal of Consumer Psychology*, *20*, 90–98. https://doi.org/10.1016/j.jcps.2009.09.003.
Jackson, D. L., Gillaspy Jr., J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694.
James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, *69*, 307–321.
Kline, R. B. (2016). *Principles and practice of structural equation modeling*, (4th ed.,). New York: The Guilford Press.
Lee, W., Lee, M.-J., & Bong, M. (2014). Testing interest and self-efficacy as predictors of academic self-regulation and achievement. *Contemporary Educational Psychology*, *39*, 86–99.
Lei, P., & Wu, Q. (2007). *Introduction to structural equation modeling: Issues and practical considerations. Educational Measurement: Issues and Practice, fall*, (pp. 33–43).
Li, X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development*, *11*, 393–404.
Liang, G. (2011). A study on effectively promoting student extracurricular English learning activities. *Chinese and Foreign Education Research*, *3*, 25–26.
MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128.
Marefat, F., & Barbari, F. (2009). The relationship between out-of-class language learning strategy use and reading comprehension ability. *Porta Linguarum*, *12*, 91–106.
Ministry of Education of the People's Republic of China. (2017). The reply to the NO. 5574 proposal submitted in the fifth session of the 12th National People's Congress. http://www.moe.gov.cn/jyb_xxgk/xxgk_jyta/jyta_jijiaosi/201712/t20171219_321937.html. Accessed 12 Feb 2020.
Ministry of Education of the People's Republic of China. (2019). The national unified syllabus of Gaokao in 2019. http://gaokao.neea.edu.cn/html1/report/19012/5951-1.htm. Accessed 24 Jan 2021.
Ministry of Education of the People's Republic of China. (2020). Making the best preparation for the 2020 GaoKao with the highest standard and the most stringent measures. http://www.gov.cn/xinwen/2020-07/02/content_5523462.html. Accessed 11 July 2020.
Oliveira, I. M., Taveira, M. C., Porfeli, E. J., & Grace, R. C. (2018). Confirmatory study of the Multidimensional Scales of Perceived Self-Efficacy with children. *Universitas Psychologica*, *17*(1), 1–12. https://doi.org/10.11144/Javeriana.upsy17-4.csms.
Pan, Y. C. (2014). Learner washback variability in standardized exit tests. *The Electronic Journal for English as a Second Language*, *18*(2), 1–30.
Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.
Qi, L. (2004). Has a high-stakes test produced the intended changes? In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods*, (pp. 171–190). New Jersey: Lawrence Erlbaum Associates.
Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*, (2nd ed.,). New Jersey: Lawrence Erlbaum Associates.
Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing*, *14*(3), 304–314.
Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behavior during test preparation. *Language Testing and Assessment*, *8*(1), 69–95.
Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231.
Shi, Y., & Jia, D. (2015). Gao kao [A documentary]. In *Zhongshi Media Corporation*. Beijing Zhongshi Beijing Film and Television Production: Company.
Stone, C. A., & Sobel, M. E. (1990). The robustness of total indirect effects in covariance structure models estimated with maximum likelihood. *Psychometrika*, *55*, 337–352.
Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment*, (3rd ed., pp. 359–372). Cham, Switzerland: Springer International Publishing AG.
Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods*, (pp. 19–36). New Jersey: Lawrence Erlbaum Associates.
Williams, J. E., & Coombs, W. T. (1996). *An analysis of the reliability and validity of Bandura's multidimensional scales of perceived self-efficacy [Paper presentation]. Annual Meeting of the American Educational Research Association*. New York: NY, United States.
Xie, Q. (2015). Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism. *System*, *50*, 56–68.
Xie, Q., & Andrews, F. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing*, *30*(1), 49–70.
Yung, Y.-F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques*, (pp. 195–226). New Jersey: Lawrence Erlbaum Associates.

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: Insights from possible self theories. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 71–89. https://doi.org/10.1080/0969594X.2012.757546.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, *29*, 663–676.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*, 51–59.

Zou, S., & Dong, M. (2014). Washback research of the recent two decades in China: Current situation and thought. *China Foreign Language*, *4*, 4–14.

Publisher's Note