Research article

Impact of learners’ perceptions of a high-stakes test on their learning motivation and learning time allotment: A study on the washback mechanism

Manxia Dong a, Xiaohua Liu b,∗

a School of Business English, Sichuan International Studies University, China
b School of Humanities and Social Science, The Chinese University of Hong Kong, Shenzhen, China

ARTICLE INFO

Keywords:
Washback mechanism
Test perceptions
Learning motivation
Learning time allotment

ABSTRACT

This study investigates the relationships between learners’ perceptions of a high-stakes test and their learning motivations and time allotment to explore the mechanism of test washback on learning. A questionnaire was administered to 3105 Chinese senior high school students. Descriptive statistics, exploratory factor analysis, standard multiple regression and structural equation modelling were performed. The study found that students’ positive test perceptions (i.e., validity and importance) better predicted their intrinsic motivations (i.e., communicative and development motivations) than their negative test perceptions (i.e., perceptions of test impact), which were found to better predict their extrinsic motivation towards external requirements. In addition, students’ perceptions of test validity and test impact both had direct effects on learning time allotment and a small indirect effect through development motivation and requirement motivation, while perceptions of test importance only indirectly influenced learning time allotment through development motivation. These findings have important implications for learning and teaching.

1. Introduction

It has been frequently found that high-stakes tests tend to exert influences on teaching and learning, and these influences are now widely known as washback in the fields of general education and applied linguistics (Alderson and Wall, 1993; Green, 2013). Since the seminal work of Alderson and Wall (1993), the number of empirical studies on washback has increased considerably (e.g., Ali and Hamid, 2020; Allen, 2016; Fan et al., 2014; Green, 2007; Sadeghi et al., 2021; Sato, 2018; Zhang and Bournot-Trites, 2021). However, most of the existing washback studies are concerned with topics such as whether a test exerts washback and what the washback looks like, while the mechanism of washback has received less attention (Cheng et al., 2004; Xie and Andrews, 2013). Xie and Andrews (2013) explained that this trend was mainly due to the methodological limitations of most washback studies, which primarily took a descriptive or qualitative approach. Qualitative methods are useful in identifying the factors affecting the nature and extent of washback; however, they are not capable of testing the specific relationships between variables. On the other hand, quantitative methods can answer questions such as in what ways and to what extent different factors that are associated with a test influence teaching and learning, which is helpful in exploring the working mechanism of washback (Xie, 2015). Thus, this study attempted to utilize quantitative methods, i.e., standard multiple regression (SMR) and structural equation modelling (SEM), to examine the relationships among test perceptions, learning motivation and time allotment with the purpose of exploring the washback mechanism of a high-stakes test.

2. Literature review

As some researchers call for more attention to washback mechanisms, in recent years, an increasing number of quantitative studies have been conducted to examine the relationships between test-takers’ test perceptions and their learning activities or performances...
(e.g., Elder et al., 2002; Fan and Ji, 2014; Xie and Andrews, 2013; Xie, 2015). For instance, through expectancy-value motivation theory (Wigfield and Cambria, 2010), Xie and Andrews (2013) investigated the relations between students’ perceptions of test design and test uses and their test preparation using SEM. In a subsequent study, Xie (2015) further examined the influences of students’ perceptions of component weighting and testing methods on their time management and approaches to test preparation. Despite their informative findings, both studies merely focused on test preparatory practices, which are only one kind of washback on learning and thus may fall short of revealing the whole picture. Dong (2020) distinguished four types of out-of-class learning activities performed by students. Employing SEM, she found that students’ perceptions of test validity, test impact and test importance influenced the four types of activities in different manners. In short, these studies on the washback mechanism have mainly focused on relationship between learners’ test perceptions and their learning practices (e.g., Dong, 2020; Xie and Andrews, 2013; Xie, 2015). Numerous studies have found that apart from learning practices, tests also influence students’ learning motivations (e.g., Allen, 2016; Dong et al., 2021; Shi, 2006). For instance, Dong et al. (2021) found that the National Matriculation English Test (NMET) in China affected the learning motivation of Chinese senior high school students and that the effects were mediated by gender, grade and proficiency level. Allen (2016) adopted a questionnaire survey and interview to investigate the washback of university English entrance exams on learning in Japan. His study revealed that test weights and content mediate students’ motivation for their English learning. However, to date, no studies have examined the statistical relations between test perceptions and learning motivations, leaving a blank regarding whether and to what extent learners’ test perceptions would influence their learning motivations. Gardner (1985) stated that “motivation to learn a second language is influenced by group related and context related attitudes, integrativeness and attitudes toward the learning situation, respectively” (p. 168). In this study, learner’ test perceptions can be regarded as test context related attitudes. Thus, based on Gardner (1985) and inspired by previous studies (e.g., Dong, 2020; Xie, 2015), this study posits that learning motivation is influenced by learners’ test perceptions, and this hypothesis will be statistically tested based on empirical data in the present study.

Learning time allotment partly reflects learners’ learning practices. Generally, the more time students invest into learning, the more efforts they make on learning practices. Learning time allotment has also been a concern of washback study on learning practices (e.g., Dong et al., 2021; Ferman, 2004; Tsagari, 2009). Some studies have found that a high-takes test would influence learning time allotment and that the influence would be mediated by students’ perceptions of test importance. For instance, based on students’ learning diaries, Tsagari (2009) found that perceived importance of the target test induced those students to spend more time on learning, though negative emotions such as boredom, fatigue and anxiety were observed among them. Zhan and Andrew (2014) also found in their qualitative studies that students’ perceptions of test importance (component weight for each section of the target test) affected students’ learning time allocation. The more important the students perceived the test to be, the more time they invested into learning.

The above studies adopted qualitative methods (diaries or interviews) to explore the relationship between students’ perceptions of test importance and their learning time allotment. In another study, Xie (2015) utilized SEM to examine the relationship between test-takers’ perceptions of test importance in terms of component weights and their time distribution. She also found that the more test-takers endorsed the importance of the adjusted weighting scheme, the more time they spent on the parts with higher weight and the less time they spent on those with lower weight. These existing studies defined test importance according to test component weights. Apart from test component weight, test importance can also be defined based on test consequences (i.e., promoting learners’ future learning). It remains unclear whether students’ perceptions of test importance in terms of test consequences would influence their learning time allotment. Meanwhile, using SEM, Dong (2020) found that learners’ perceptions of test validity and test impact influenced their learning activities in different ways and to different degrees, but whether perceptions of test validity and test impact would influence learning time allotment is unknown.

In light of the above research gaps in washback, this study surveyed a large sample of Chinese senior high students to examine the relationship between their test perceptions, learning motivation and learning time allotment with the purpose of exploring how a high-stakes test influences their learning (or washback working mechanism). Such exploration of test washback mechanism will be of interest to learners and teachers because it can provide useful feedback and important implications for improving learning and teaching. This study aimed to address the following two research questions.

RQ1. Do learners’ perceptions of test validity, impact and importance influence learning motivation and learning time allotment simultaneously?

RQ2. If so, in what ways and to what extent do test perceptions influence them?

3. Research context

This study was conducted within the secondary education context of China. Tests have played a significant role in Chinese education system. Qi (2007) even claimed that the Chinese educational system is an examination-oriented system. Among all large scale high-stakes tests, the National Matriculation Test (NMT) has been regarded as the most influential, significant and competitive high-stakes test in China. The NMT score is not only used to make a university or college admission decision, but also used by schools, parents, and students to evaluate teachers and by society to evaluate schools (Qi, 2004).

The NMT battery usually includes five or six subjects, among which the English test, namely the NMET, is one of the three compulsory tests for all the candidates. The NMET was developed by the National Education Examinations Authority and is administered to approximately 10 million high school graduates annually across Mainland China. Apart from being used to make the university admission decision together with other subject tests, the NMET, as a high stakes test, is also designed to generate a positive washback effect on secondary English teaching and learning (Qi, 2004). Although some researchers have conducted washback studies on the NMET (e.g., Li, 1990; Qi, 2007), very few studies have explored how NMET washback works. Thus, the NMET is an excellent case for exploring this issue in this study. The NMET is a paper-and-pencil test with a total mark of 150, routinely consisting of four parts: listening (20%), reading (26.7%), language usage (30%) and writing (23.4%). Listening, reading and language usage employ objective tasks (i.e., multiple-choice questions) that are rated by computer, while writing adopts subjective tasks and is marked by human raters.

4. Research method

To answer the two research questions, this study adopted a quantitative approach through questionnaire survey. Fig. 1 offers an overview of the research process. In the remaining parts of this section, details regarding instrument development, participants selection, and data collection and analyses will be presented.

4.1. Participants and data collection

In China, secondary education schools are divided into different categories according to their prestige and educational accomplishments, including top schools, key schools and ordinary schools. Key schools usually include city/province key schools and district/county key schools. Ordinary schools include city/province ordinary schools, district/county ordinary schools and town ordinary schools. To obtain
a more representative sample, we selected six different types of high schools and included each grade from the sampled schools. Given that the classes in each grade in nearly all schools were divided into different tiers (usually two or three tiers) according to the comprehensive academic performance of the students to cultivate the targeted candidates for different tiers of universities, the number of tiers in each grade was taken into account in the selection of the sampled classes. Finally, 56 classes with a total of 3278 senior high students spanning three grades from six tiers of high schools located in a southwestern city of China were recruited. The six sampled schools included a city top high school (N = 429, 13.8%), a city key high school (N = 716, 23.1%), a district key high school (N = 551, 17.7%), a district ordinary high school (N = 642, 20.7%), a county ordinary high school (N = 642, 20.7%), a county ordinary high school (N = 303, 9.8%) and a town ordinary high school (N = 464, 15.0%).

The survey was conducted at the end of the academic year (late May and early June). Before that, ethics approval was obtained from Shanghai International Studies University. Prior to collecting data, all the participants were informed of the purpose and the voluntary nature of the survey, and then signed a consent form for their participation. All hard-copy questionnaires were then completed in class by students with the assistance of their English teachers, who were provided with an administration manual and designated to be in charge of the administration of the questionnaire survey. Once completed, the questionnaires were immediately collected and mailed back to the first author by the designated teachers. Among the 3278 returned questionnaires, 173 were found to have at least 10% incomplete questions or obvious response patterns and were thus considered invalid. Of the 3105 questionnaires, 38.6% were from Year 1 (N = 1199), 35.4% were from Year 2 (N = 1098), and 26% were from Year 3 (N = 808). Additionally, there were 1402 male respondents (45.2%) and 1703 female respondents (54.8%). The average age of the participants was 17 (SD = 1.043).

4.2. Instruments

The student questionnaire was developed from a group interview with six senior high students as well as references to previous studies (Gao et al., 2003; Qi, 2004). It contained questions asking about students’ demographic information (i.e., their gender, age, grade), their perceptions of aspects of the NMET (i.e., the test perception scale), their learning motivation (i.e., the learning motivation scale), and a question asking about their daily time spent on English learning outside class. Questions from the two scales were rated on a 5-point Likert-scale of agreement, ranging from 1 (“disagree”) to 5 (“agree”) (see the Appendix for an English translation of the questionnaire). The questionnaire underwent careful scrutiny by the researchers, language assessment specialists, high school English teachers and students to identify points of ambiguity, irrelevance and redundancy in the items to ensure the accuracy, appropriateness, and clarity of the wording for each item. The draft questionnaire was piloted with 179 senior high school students, including 51 Year 1 students, 66 Year 2 students, and 62 Year 3 students. Generally, the analysis of the pilot data demonstrated that the questionnaire had a satisfactory reliability (α = .846).

A scale-level exploratory factor analysis (EFA) was performed on the full sample data in the main study to examine the construct validity of the questionnaire. In EFA, the factors were extracted by principal component analysis (varimax rotation) and based on eigenvalues greater than 1.

The EFA of the test perception scale revealed three underlying factors, including perceptions of test validity (Pvali), perceptions of test impact (Pimpia) and perceptions of test importance (Pimpio). Pvali had three items, which were concerned with students’ perceptions of the validity of NMET (e.g., “NMET examines my senior high English learning”). Pimpia had 4 items that addressed students’ perceptions of NMET influence (e.g., “NMET pressures me to review test-taking skills”). Perception of test importance (3 items) concerned learners’ perceptions of the importance of the NMET (e.g. “My NMET score gives me a feeling of pride”).

The EFA of the learning motivation scale also revealed three factors, which cumulatively explained 65.8% of the total variance. The three factors represented three types of learning motivation, including communication motivation (CM), development motivation (DM), and requirement motivation (RM). CM had 4 items that described English learning for communicative purposes (e.g., “I learn English because I hope to communicate with other people in English”). DM also had 4 items, which represented students’ English learning for future study and career development (e.g., “I learn English because it can help me to have a better job in the future”). RM had three items, which were related to English learning for the purpose of meeting external requirements (e.g., “I learn English because English is a compulsory course”).

Students’ English learning time outside class was measured by one item with a 5-point scale ranging from 1 (“None”) to 5 (“91-120 minutes”). The higher the scores, the more time was invested by students in English learning.

4.3. Data analysis

In this study, SPSS 21.0 and AMOS 21.0 were used to analyze the data. Before the data analysis, all the data went through several checks to identify any errors, missing values and outliers. In the 3105 valid cases, only 3 questionnaire copies had missing values (fewer than 2 missing values for each copy). The missing values were handled by imputing the mean value of the item, and outliers were identified by examining the stem-and-leaf plot for each item (Tabachnick and Fidell, 2007). The skewness and kurtosis statistics of all items were then examined to evaluate if normality assumptions were met.

Based on the literature review, this study hypothesized the general relationship between learners’ test perceptions, learning motivation
Table 1. Descriptive statistics and correlation analysis of the examined variables (N = 1553).

| Variables | Descriptives | Correlation Coefficient |
|-----------|--------------|-------------------------|
|           | M   | SD  | Pvali | Pimp | Pimpo | CM     | RM     | DM     | Letime |
| Pvali     | 3.23| 1.143| —     | .140** | .491* | .258** | .069** | .389*  | .255** |
| Pimp      | 3.43| .981 | —     | .296** | .060* | .237** | .073** | .152** |        |
| Pimpo     | 3.29| 1.140| —     | .218** | .229**| .375** | .218** |        |        |
| CM        | 3.43| 1.102| —     | -.128**| .536**| .176** |        |        |        |
| RM        | 3.71| 1.103| —     | .014  | -.002 |        |        |        |        |
| DM        | 4.10| .954 | —     | .269**|        |        |        |        |        |
| Letime    | 2.42| 1.026| —     |        |        |        |        |        |        |

Notes: Pvali = perceptions of test validity, Pimp = perceptions of test importance, Pimpo = perceptions of test motivation, CM = communication motivation, DM = development motivation, RM = requirement motivation, Letime = learning time; * p < .05; ** p < .01; *** p < .001.

Table 2. Standard multiple regression statistics (N = 1553).

| Variables | R   | R²  | Adjust R² | F (3,1549) (6,1546) | β  | t   | p    | Tolerance | VIF |
|-----------|-----|-----|-----------|---------------------|----|-----|------|------------|-----|
| IV        | .278| .077| .076      | 43.333              |    |     |      |            |     |
| Pvali     | .198| 7.080|.000  | .759 | 1.318 |
| Pimp      | -.003| -.125| .900  | .913 | 1.096 |
| Pimpo     | .122| 4.185|.000  | .706 | 1.416 |
| DV        | .294| .086| .085     | 48.803              |    |     |      |            |     |
| Pvali     | -.056| -.200| .046  | .759 | 1.318 |
| Pimp      | .186| 7.304|.000  | .913 | 1.096 |
| Pimpo     | .201| 6.969|.000  | .706 | 1.416 |
| DV        | .444| .197| .196     | 126.947           |    |     |      |            |     |
| Pvali     | .271| 10.350|.000  | .759 | 1.318 |
| Pimp      | -.040| -.1674| .094  | .913 | 1.096 |
| Pimpo     | .253| 9.353|.000  | .706 | 1.416 |
| DV        | .341| .116| .113     | 33.861              |    |     |      |            |     |
| Pvali     | .142| 4.979|.000  | .707 | 1.415 |
| Pimp      | .114| 4.452|.000  | .879 | 1.138 |
| Pimpo     | .057| 1.899|.058  | .645 | 1.550 |
| CM        | .021| .732| .464     | .688 | 1.454 |
| RM        | .174| -.2026|.043  | .882 | 1.134 |
| DM        | -.052| 5.708|.000  | .617 | 1.621 |

Notes: Pvali = perceptions of test validity, Pimp = perceptions of test importance, Pimpo = perceptions of test motivation, CM = communication motivation, DM = development motivation, RM = requirement motivation, Letime = learning time; * p < .05; ** p < .01; *** p < .001.

and learning time allotment. However, this hypothesis did not suggest the relationships between specific variables. Therefore, this study attempted to conduct a cross-validation by randomly splitting the total sample (N = 3105) into halves. We used the first half of the sample (N = 1553), whose cases were randomly selected through SPSS, to conduct a standard multiple regression (SMR) to explore the impact of learners’ test perceptions on their learning motivation and learning time allotment. Based on the results of SMR analysis, we then conducted SEM using AMOS 21.0 with the second half of the sample (N = 1552) to confirm the exact relationships between those variables. To evaluate model fit, indices, such as SRMR, chi-square, CFI, TLI and RMSEA with its confidence intervals were consulted. Generally, the smaller the chi-square value is, the better the data fits the model. SRMR and RMSEA ≤ 0.08 and CFI and TLI ≥ 0.9, are considered good fits (Bentler, 1990; Hu and Bentler, 1999).

5. Results

5.1. Standard multiple regression analysis

Before the SMR analysis, descriptive statistics and Pearson bivariate correlations at the factor level needed to be calculated. Thus, we conducted an EFA to explore the factor structure. The results showed that the factor structure in the first half of the sample was consistent with that of the whole sample, which also confirmed the construct validity of the questionnaire. Descriptive statistics and Pearson bivariate correlations at the factor level on the first half of the sample are displayed in Table 1.

Table 1 showed that Pimpo had largest mean (M = 3.43, SD = .981), followed by Pimpo (M = 3.29, SD = 1.140) and Pvali (M = 3.23, SD = 1.143), suggesting that students had the highest level of agreement regarding test impacts on them. Among the three motivation factors, development motivation had the largest mean (M = 4.10, SD = .954), followed by requirement motivation (M = 3.71, SD = 1.103), and communicative motivation had the smallest mean value (M = 3.43, SD = 1.102). The mean value of learning time allotment was 2.42 (SD = 1.026), indicating that students spent from half an hour to one hour on learning outside class each day. Correlation analysis at the factor level of the perceptions scale, learning motivation scale, and learning time allotment scale showed that no factor correlation exceeded 0.80, indicating good discriminant validity across the factors in the scales (Brown, 2015). Table 1 also shows that only the RM variable was not significantly related to DM and Letime, while the other variables were all significantly correlated with each other.

The two diagnostic indices of tolerance values (> .20) and VIF (< 5.0) were within satisfactory ranges, indicating that variables within this data set did not suffer from the multicollinearity problem (Xie, 2013). Thus, these variables could be used in the SMR. The results of the SMR are presented in Table 2.

Table 2 showed that Pvali, Pimp and Pimpo had small but significant contributions to students’ CM (R² = .077, p < .05), explaining 7.7% of the variance. Among the three variables, Pvali and Pimpo had small but significant effects (p < .001) on CM (β = .198, β = .122), while Pimpo had a very small negative nonsignificant effect on CM (β = -.003, p = .900). Pimpo and Pimpo had significant effects (p < .001) on RM (β = .186, β = .201), while Pimpo had a significantly negative but
Table 3. Descriptive statistics on the second half of sample (N = 1552).

| Dimensions                      | Factor | Variables | Min | Max | M   | SD  | Skewness | Kurtosis |
|---------------------------------|--------|-----------|-----|-----|-----|-----|----------|----------|
| Perceptions of a test           | Pvali  | Pvali1    | 1 5 | 3.33| 3.19| 1.19| -0.056   | -1.182   |
| (M = 3.26, Pvali2)              |        |           |     |     |     |     |          |          |
| SD (1.165)                      | Pvali3  | 1 5       | 3.29| 3.23| 0.82| 0.28| -0.957   | -0.986   |
| Pimpa                           | Pimpa1  | 1 5       | 3.06| 1.29| -0.028| -0.968 |
| (M = 3.39, Pimpa2)              |        |           |     |     |     |     |          |          |
| SD (0.989)                      | Pimpa3  | 1 5       | 3.60| 1.261| -0.630| -0.604 |
| Pimpa4                           | 1 5    | 3.26      | 1.423| -0.251| -1.205 |
| SD (1.131)                      | Pimpa5  | 1 5       | 3.45| 1.340| -0.528| -0.850 |
| Pimpa6                           | 1 5    | 2.98      | 1.418| -1.282| -1.131 |
| Learning motivation             | CM     | CM1       | 5 1| 3.61| 3.84| -0.639| -0.878   |
| (M = 3.48, CM2)                 |        |           |     |     |     |     |          |          |
| SD (1.096)                      | CM3    | 1 5       | 3.82| 3.92| -0.233|
| CM4                            |        |           |     |     |     |     |          |          |
| DM                              | DM1    | 1 5       | 3.47| 3.48| -0.911|
| (M = 4.12, DM2)                 |        |           |     |     |     |     |          |          |
| SD (0.926)                      | DM3    | 1 5       | 4.05| 1.176| -1.396| -0.548 |
| DM4                            |        |           |     |     |     |     |          |          |
| RM                              | RM1    | 1 5       | 3.59| 1.433| -0.657| -0.930 |
| (M = 3.65, RM2)                 |        |           |     |     |     |     |          |          |
| SD (1.131)                      | RM3    | 1 5       | 3.83| 3.82| -0.745|
| Letime                          |        |           |     |     |     |     |          |          |
| Note: Pvali = perceptions of test validity, Pimpa = perceptions of test impact, Pimpo = perceptions of test importance, CM = communication motivation, DM = development motivation, RM = requirement motivation, Letime = learning time.

Table 4. Model fit indexes for the hypothesized model.

| $\chi^2$ | df | CFI | TLI | SRMR | RMSEA 90% CI |
|----------|----|-----|-----|------|-------------|
| Acceptable fit | ≥.9|≥.9 |<.08|<.05 ≤ x ≤ .08 |
| Hypothesized model (Fig. 1) | 1341.3|193 |.908|.890| .064,.062[.059,.065] |
| Structural model without correlating error terms | 1830.1|196 |.869|.846| .095,.073[.070,.076] |

Note: $\chi^2$ = chi-square, df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis index, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation.

Figure 2. A graphic representation of this final model with standardized coefficients, which are the estimates of the impact of perceived test validity, test impact, and impact on English learning motivation and learning time allotment. The structural model, the error terms of the variables relating to learning motivation were fixed to be correlated due to the results of correlation analysis of CM, DM, and RM, as well as the method effect (Xie, 2015). That is, the three latent variables of learning motivation shared similar wording and were listed together for the learners. By fixing these error correlations among the latent variables of learning motivation, the structural model recognizes the possibility that learners' learning motivation may well be influenced by the factors not included in the model (Byrne, 2011). To warrant the appropriateness of the error terms, we made model comparisons between the final model and a model without the correlated error terms. The model fit indices displayed in Table 4 show that the hypothesized model was significantly better than a model without the correlated error terms ($\Delta \chi^2 = 691.9, df = 3, p < .001$), which provides further support for correlating these error terms.

5.2. Structure equation modelling

The descriptive statistics of the second half of the sample (N = 1552) at the factor and item levels are displayed in Table 3. The results of most statistics of skewness and kurtosis values showed that they were within or very close to the satisfactory range (±1.0). Thus, these variables were regarded as having the normality assumptions.

Learners' perceptions of the NMET included perception of test validity (Pvali) (M = 3.26, SD = 1.165), perception of test impact (Pimpa) (M = 3.39, SD = .989) and perception of test importance (Pimpo) (M = 3.28, SD = 1.131). The results showed that the mean values of perceptions were above three on a 5-point scale, indicating that students tended to agree with the test validity, impact and importance of the NMET. Among the three types of perceptions, students were more likely to agree with test impact than test validity and importance.

English learning motivation involved communicative motivation (CM) (M = 3.48, SD = 1.096), development motivation (DM) (M = 4.12, SD = .926) and requirement motivation (RM) (M = 3.65, SD = 1.131). The mean values of the three types of learning motivation on a 5-point scale all exceeded the midpoint, 3.0, suggesting a generally favourable disposition towards English learning motivation. Students tended to have stronger development motivation (M = 4.11) than communicative motivation or requirement motivation. The mean value of learning time investment was 2.44 (SD = 1.050) on a five-point scale, indicating that students spent half an hour to one hour on English learning outside class each day.

Fig. 2 is a graphic representation of this final model with standardized coefficients, which are the estimates of the impact of perceived test validity, test impact, and impact on English learning motivation and learning time allotment. In the structural model, the error terms of the variables relating to learning motivation were fixed to be correlated due to the results of correlation analysis of CM, DM and RM, as well as the method effect (Xie, 2015). That is, the three latent variables of learning motivation shared similar wording and were listed together for the learners. By fixing these error correlations among the latent variables of learning motivation, the structural model recognizes the possibility that learners' learning motivation may well be influenced by the factors not included in the model (Byrne, 2011). To warrant the appropriateness of the error terms, we made model comparisons between the final model and a model without the correlated error terms. The model fit indices displayed in Table 4 show that the hypothesized model was significantly better than a model without the correlated error terms ($\Delta \chi^2 = 691.9, df = 3, p < .001$), which provides further support for correlating these error terms.

The model fit indices of the hypothesized model were as follows: $\chi^2 = 1138.2, df = 193, CFI = .908, TFI = .890, SRMR = .064, RMSEA = .062$ [90% CI = .059, .065]. Due to the extreme sensitivity of the chi-square statistic to sample size, we had to consider the other indices. For this model, CFI = .908 and TFI = .890, which were greater than or very close to .90, and the SRMR and RMSEA values were less than .080 and thus were within acceptable ranges. Given all of the indices, this model achieved a good fit and explained the underlying pattern of the impact of learners' perceptions on learning motivation and learning time allotment simultaneously. We also found that nearly all the results...
were consistent with the findings in the SMR, except for the two paths (Pvali → RM) ($\beta = -0.04, p = .433$) and RM → Letime ($\beta = -0.05, p = .078$). We found that the regression coefficients in the SMR analysis were very close to the nonsignificant level ($\beta = -0.056, p = .043; \beta = .174, p = .043$), which might explain the small differences in terms of the results between SMR analysis and SEM.

Table 5 presents the specific effects of learners' test perceptions on learning motivation and learning time allotment. The results indicated that learners' favourable perception of test validity had a positive and significant influence ($p < .001$) on their communicative motivation (CM) ($\beta = .11$) and development motivation (DM) ($\beta = .15$) but exhibited a statistically nonsignificant negative effect on requirement motivation (RM) ($\beta = -0.04, p = .433$). The perception of impact (Pimpa) exerted positive and significant influences only on requirement motivation ($\beta = .20, p < .001$). The model also showed that the perception of importance (Pimpo) had positive and significant effects ($p < .001$) on CM ($\beta = .27$), RM ($\beta = .10$) and DM ($\beta = .42$).

The results also showed that Pvali exerted both a direct effect on Letime ($\beta = .16, p < .001$), and a small but statistically significant indirect effect ($\beta = .02, p < .01$) via the mediation of DM but had basically no indirect effect via the mediation of RM ($\beta = .02, p = .544$). The total effect of Pvali on Letime was statistically significant ($\beta = .14, p < .001$). Learners' perceptions of test impact also exerted both a direct effect on Letime ($\beta = .14, p < .001$), and a small but nonsignificant indirect effect ($\beta = -0.01, p = .091$). The total effect of Pimpa on Letime was statistically significant ($\beta = .13, p < .001$). Meanwhile, Pimpo did not exert a direct effect on Letime but had a small significant indirect effect on Letime via the mediation of DM and had no significant effect on Letime via the mediation of RM ($\beta = -0.01, p = .178$). The total effect of Pimpa on Letime was statistically significant, although the effect was small ($\beta = .04, p < .001$).

6. Discussion

6.1. Learners' test perceptions influencing learning motivation

In this study, learning motivation included communicative motivation, development motivation and requirement motivation. Commu-
Communicative motivation refers to learners’ English learning for communicative purposes. Learners with communicative motivation engage in learning activities with stronger learning desire, interest and joy, which meets the aims of language learning. Development motivation is associated with learners’ future learning and career development, which is helpful in arousing learners’ learning interest and satisfaction. Requirement motivation refers to learning primarily for external requirements, e.g., learning for the test, learning driven by curriculum arrangement and other individuals’ expectation and curriculum arrangements. Dörnyei and Ushioda (2011) divided motivation into intrinsic versus extrinsic motivation and proposed that “intrinsic motivation (IM) deals with behaviour performed for its own sake in order to experience pleasure and satisfaction, such as the joys of doing a particular activity or satisfying one’s curiosity, while extrinsic motivation involves performing a behaviour as a means to some separable end, such as receiving an extrinsic reward (e.g. good grades) or avoiding punishment” (p. 23). According to Dörnyei and Ushioda (2011), communicative motivation and development motivation in this study tended to be associated with learners’ intrinsic motivation, and requirement motivation was related to extrinsic motivation.

This study found that learners’ favourable perceptions of test validity positively and significantly influenced communicative motivation (CM) ($\beta = .11$) and development motivation (DM) ($\beta = .15$) but had a negative nonsignificant effect on requirement motivation (RM) ($\beta = -.04$, $p = .433$). The results indicate that those learners who held favourable perceptions of test validity tended to have stronger intrinsic motivation but had very weak extrinsic motivation. The result is plausible in the washback context because only when students recognize the validity of a test in examining their learning achievement or ability can their strong intrinsic desire for learning be motivated; therefore, they make more efforts to engage in learning practices (i.e., regular learning and test preparation) to meet the demands or goals set by the test. In contrast, if students perceive a test to be invalid or perceive the test quality to be poor in assessing their proficiency, they naturally do not show interest or desire to learn, hence, their intrinsic learning motivation will be reduced or even stay deactivated.

In this study, test impacts included four items, which depicted the effects of the NMET on learners’ English learning, including “the NMET orients my English learning to the test”, “The NMET pressures me to review test-taking skills” and “The NMET preparation takes time away from regular English learning”. Some previous studies have documented negative test impacts, such as taking time away from regular teaching and learning due to test preparation, narrowing curriculum content, memorizing model essays by rote, excessively practising test items, focusing on test-taking skills (Cooley, 1991; Corbett and Wilson, 1991; Stecher et al., 2004). According to these descriptions, the test impact surveyed in the present study can also be categorized as negative impact. The current study also found that the more learners agreed that they received negative impact from the test, the more likely for them to endorse requirement motivation (extrinsic learning motivation) ($\beta = .20$). This is probably because those who perceived negative influence of the test on them tended to engage in more test preparation or test-oriented learning, which has been supported by previous studies (e.g., Xie, 2015; Dong, 2020). Meanwhile, their engagement in test preparation practices in turn seems to have boosted their extrinsic motivation and prompted them to pursue external rewards (high scores in the NMET in this case) (Dörnyei and Ushioda, 2011).

The findings also suggest that learners’ perception of test importance (Pimpo) significantly influenced all three types of learning motivation. Specifically, learners’ perceived test importance exerted medium-sized effects on development motivation ($\beta = .42$) and communicative learning ($\beta = .27$) and a small effect on requirement motivation ($\beta = .10$). These results indicate that learners’ favourable perceptions of test importance are likely to spark both intrinsic and extrinsic motivation. This echoes Credé and Phillips’s (2011) finding that the perceived importance of tasks (task value) is generally found to be positively correlated with both intrinsic and extrinsic goal orientations/motivation. A possible reason is that those who realized the importance of the NMET for their future study, career development and interpersonal communication had a stronger learning desire for promoting their future development and communication, thereby stimulating their stronger intrinsic motive.

### 6.2. Learners’ test perceptions influencing learning time allotment

The study found that learners’ favourable perception of test validity significantly directly influenced their learning time allotment. The more positive their attitudes towards test validity, the more time students will spend on learning. Previous studies have found that learners’ favourable perceptions of test validity will prompt them to engage in more frequent learning practices (e.g., Dong, 2020; Xie, 2015), which means more time being invested into learning. Thus, we deem that this result is basically consistent with previous findings. The results also show that development motivation statistically significantly mediated the influence of the perception of test validity on learning time allotment. This means that when students approve of test validity, they are more concerned about their future development and will invest more time into learning in order to achieve their future development goals.

Learners’ perceptions of test impact also directly affected their learning time allotment. This result indicates that when students endorsed the impact of a test on them, they would spend more time on English learning. This was primarily because learners’ perceptions of test impact positively predicted more frequent learning practices, which has been verified in previous studies in the same context (e.g., Dong, 2020).
This study found that learners’ perceptions of test importance did not exert direct effects on learning time allotment but had only indirect significant effects via the mediation of development motivation. This result suggests that development motivation could play an important role in the investment of learning time. That is, if students do not have a serious concern about their future development, even if they recognize the test importance to their learning and life, they will still not spend their time on learning. This finding seems to be slightly inconsistent with those in previous studies (Tsagari, 2009; Xie, 2015; Zhan and Andrew, 2014), which showed that testakers’ perceptions of test importance positively predicted their test preparation time. This inconsistency might be due to the following reasons. First, in previous studies, test importance was defined according to subtests’ component weights. The larger weights a subcomponent has, the more important it is perceived. In our study, test importance was more related to the whole test and its test consequences. For instance, the following items were used: “My NMET score gives me a feeling of pride”, “My NMET score is important to my future English learning”, and “My NMET score enhances my self-confidence in English learning”. Such difference in construct definition may cause the inconsistency in finding. In addition, previous studies have primarily adopted qualitative method (i.e., learning diaries and interviews) to explore the influence of learners’ perceptions of test importance on their learning time. Their results were tentative and exploratory due to the small sample sizes. Moreover, previous studies did not test the mediating effect of learning motivation between learners’ perceptions of test importance and learning time allotment. Through testing the mediating effect of learning motivation, the present study provides a new insight into the relationship between the two variables.

7. Conclusion

The present study examined the relationships between learners’ test perceptions and learning motivation and learning time allotment using SMR and SEM, aiming to reveal how a test influences their learning motivation and learning time allotment. The results showed that learners’ endorsement of test validity, test impact and test importance affected their learning motivation in different ways and to different degrees. Specifically, learners’ endorsement of test validity and test importance were more related to their stronger communicative motivation and development motivation, whereas their endorsement of negative test impact only significantly affected requirement motivation. The study determined that learners’ positive test perceptions predicted stronger intrinsic motivation; conversely, their negative perceptions of test impact were more associated with extrinsic motivation. The study also found that learners’ test perceptions significantly predicted their learning time allotments. The perceptions of test validity and test impact had a direct effect and a small indirect effect through development motivation and requirement motivation, while perceptions of test importance significantly influence learning time allotment through development motivation.

The exploration of the relationship between learners’ test perceptions, learning motivation and learning time allotment is beneficial to enriching and deepening our understanding of the mechanism of washback on learning. This study has some important implications for improving English learning and teaching. This study found that learners’ test perceptions influence their learning motivation and learning time allotment. Thus, in regular teaching and learning, teachers and students should attach importance to stakeholders’ test perceptions. Specifically, it is suggested that learners hold positive attitudes towards tests, which are conducive to arousing stronger intrinsic motivation and investing more learning time, which in turn helps enhance the positive test effect and reduce negative effect. In the EFL context, teachers are the most direct individuals influencing learners’ test perceptions, learning motivation and learning practices. Thus, it is also suggested that teachers themselves promote the proper perspective of a test and guide their students to maintain positive attitudes towards a test through a variety of platforms and ways to increase their intrinsic learning motivation. In addition, we should take measures to nurture students’ intrinsic learning motivation, which prompts students to invest learning time and focus on deep and meaningful learning (Wigfield and Cambria, 2010).

The major limitation of this study is that in the main study only self-reported questionnaire data were collected, which reduces the robustness of the research results to some extent due to lack of evidence from other sources, such as class observation and interviews data. As Alderson and Wall (1993) suggested, whether what teachers and learners say is reflected in their behaviour needs support from observational data. Moreover, the learning time allotment scale only had one item, and thus, it may suffer from low reliability due to measurement error and the inability to capture the full spectrum of the construct. In addition, it is difficult to tease out the directionality of the relations and/or remove the influence of unmeasured confounding variables due to lack of experimental manipulation and longitudinal relations. Thus, the mediation tests in the model should be interpreted with caution. Although we considered the variables of school tiers, grade and gender in selecting the participants, these variables were not included in the data analysis. Future studies can conduct multi-group SEM to examine the stability of the model across groups of participants, which is helpful to provide a more comprehensive understanding of washback mechanism. In addition, future research can consider conducting a longitudinal study to track the dynamic nature of test washback over time.

Declarations

Author contribution statement

Manxia Dong: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Xiaohua Liu: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by National Social Science Fund of China [18XYY014], Sichuan International Studies University [sisu201716].

Data availability statement

Data will be made available on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Appendix A. English translation of questionnaire

Part one: personal data
1. Your Gender: A. Female B. Male
2. Your Grade: A. Grade 1 B. Grade 2 C. Grade 3
3. Your Age: ___ (fill in the blank).

Part two: learning time

Please read the following statement and circle the one that seems most appropriate to you:

The time you spend on English learning outside class each day:
1 = none 2 = 1–30 minutes 3 = 31–60 minutes 4 = 61–90 minutes 5 = 91–120 minutes

Part three: perceptions of NMET
Please grade the following statements of senior high students’ opinions regarding NMET using a 5-point scale format: 1 = disagree, 2 = somewhat disagree, 3 = undecided, 4 = somewhat agree, 5 = agree.

**Pval1** NMET reflects my English level scientifically and objectively. 1 2 3 4 5

**Pval2** NMET examines my senior high English learning. 1 2 3 4 5

**Pval3** NMET facilitates my setting a clear goal with respect to learning English. 1 2 3 4 5

**Pimp1** NMET orients my English learning to the test. 1 2 3 4 5

**Pimp2** NMET preparation takes time away from regular English learning. 1 2 3 4 5

**Pimp3** NMET pressures me to review test-taking skills. 1 2 3 4 5

**Pimp4** NMET content focuses my English learning. 1 2 3 4 5

**Pimp5** My NMET score gives me a feeling of pride. 1 2 3 4 5

**Pimp6** My NMET score is important to my future English learning. 1 2 3 4 5

**Pimp7** My NMET score enhances my self-confidence in English learning. 1 2 3 4 5

Part four: English learning motivation

Please grade the following statements concerning senior high students’ opinions about the National Matriculation English Test (NMET) using the following 5-point scale: 1 = disagree, 2 = somewhat disagree, 3 = undecided, 4 = somewhat agree, 5 = agree.

**CM1** I learn English because I hope to communicate with other people in English. 1 2 3 4 5

**CM2** I learn English because I hope to understand American and English culture and customs. 1 2 3 4 5

**CM3** I learn English because I hope to study abroad in the future. 1 2 3 4 5

**CM4** I learn English because I hope to make foreign friends. 1 2 3 4 5

**DM1** I learn English because it can help me to develop in the future. 1 2 3 4 5

**DM2** I learn English because it can help me to have a better job in the future. 1 2 3 4 5

**DM3** I learn English because it can improve my ability to meet social needs. 1 2 3 4 5

**DM4** I learn English because it can broaden my view and enrich my knowledge. 1 2 3 4 5

**RM1** I learn English because it meets my parents’ expectations. 1 2 3 4 5

**RM2** I learn English because English is a compulsory course. 1 2 3 4 5

**RM3** I learn English because English is compulsory on the NMET. 1 2 3 4 5

References

Byrne, R.M., 2011. Structural Equation Modelling with Mplus: Basic Concepts, Applica-
tions, and Programming. Routledge.

Cheng, L.Y., Watanabe, Y., Curtis, A. (Eds.), 2004. Washback in Language Testing Re-
search: Contexts and Methods. Lawrence Erlbaum.

Cooley, W.W., 1991. State-wide student assessment. Educ. Meas., issues Pract. 10 (4), 3–6.

Corbett, H.D., Wilson, B.L., 1991. Testing, Reform, and Rebellion. Ablex.

Credé, M., Phillips, L.A., 2011. A meta-analytic review of the motivated strategies for learning questionnaire. Learn. Individ. Differ. 21 (4), 337–346.

Dong, M.X., 2020. Structural relationship between learners’ perceptions of a test, learning, and learning outcomes: a study on the washback mechanism of a high-stakes test. Stud. Educ. Eval. 64, 1–11.

Dong, M.X., Fan, J., Xu, J., 2021. Differential washback effects of a high-stakes test on students’ English learning process: evidence from a large-scale stratified survey in China. Asia Pac. J. Educ., 1–18.

Dörnyei, Z., Ushioda, E., 2011. Teaching and Researching: Motivation, 2nd ed. Pearson Education, Limited.

Elder, C., Iwashita, N., McNamara, T., 2002. Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? Lang. Test. 19 (4), 347–386.

Fan, J., Ji, F., 2014. ‘Test candidates’ attitudes and their test performance: the case of the Pudan English test. In: University of Sydney Papers in TESOL, vol. 9, pp. 1–35.

Fan, J., Ji, P., Song, X., 2014. Washback of university-based English language tests on students’ learning: a case study. Asian J. Appl. Linguist. 1 (2), 178–191.

Ferman, I., 2004. The washback of an EFL national oral matriculation test on teaching and learning, In: Cheng, L., Watanabe, Y., Curtis, A. (Eds.), Washback in Language Testing: Research Methods and Contexts. Lawrence Erlbaum, pp. 190–210.

Gao, Y.H., Zhao, Y., Cheng, Y., Zhou, Y., 2003. The relationship between types of English learning motivation and motivational intensity: a quantitative investigation on Chinese college undergraduates. Foreign Lang. Res. 2003 (1), 60–64.

Gardner, R.C., 1985. Social Psychology and Second Language Learning: The Role of Atti-
tudes and Motivation. Edward Arnold.

Green, A., 2007. Washback to learning outcomes: a comparative study of IELTS prepa-
ation and university pre-sessional language courses. Assess. Educ. Pract. Policy Pract. 14 (1), 75–97.

Green, A., 2013. Washback in language assessment. Int. J. Eng. Stud. 13 (2), 29–31.

Hu, L.T., Bientler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analy-
sis: conventional criteria versus new alternatives. Struct. Equ. Model. Multidiscip. J. 6 (1), 1–55.

Li, X., 1990. How powerful can a language test be? The MET in China. J. Multiling.

Multicult. Devi. 11 (5), 393–404.

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., 2007. Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. Assess. Educ. 14 (1), 51–74.

Sadeghi, K., Ballardj, A., Medea, E., 2021. The washback effect of TOEFL iBT and a local English Proficiency Exam on students’ motivation, autonomy and language learning strategies. Heliyon 7, e08135.

Sato, T., 2018. The impact of the test of English for academic purposes (TEAP) on Japanese students’ English learning. JACET J. 62, 89–107.

Shih, C., 2006. Perceptions of the general English proficiency test and its washback: a case study at two Taiwanese technological institutes. Unpublished doctoral dissertation. Department of Curriculum, Teaching and Learning, University of Toronto.

Stecher, B., Chun, T., Barron, S., 2004. The effects of assessment-driven reform on the teaching of writing in Washington State. In: Cheng, L., Watanabe, Y., Curtis, A. (Eds.), Washback in Language Testing: Research Contexts and Methods. Lawrence Erlbaum, pp. 53–71.

Tabachnick, B.G., Fidell, L.S., 2007. Using Multivariate Statistics, 5th ed. Pearson, Allyn & Bacon.

Taagari, D., 2009. Revisiting the concept of test washback: investigating FCE in Greek language schools. In: Cambridge ESOL: Research Notes, vol. 2009(35), pp. 5–10.

Wipfler, A., Cambria, J., 2010. Students’ achievement values, goal orientations, and in-
terest: definitions, development, and relations to achievement outcomes. Dev. Rev. 30 (1), 1–35.

Xie, Q., 2013. Does test preparation work? Implications for score validity. Lang. Assess. Q. 10 (2), 196–218.

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, S., 2013. Do test design and uses influence test preparation: testing a model of washback with structural equation modeling. Lang. Test. 30 (1), 49–70.

Zhan, Y., Andrew, S., 2014. Washback effects from a high-stakes examination on out-of-
class English learning: insights from possible self theories. Assess. Educ. Pract. Policy Pract. 21 (1), 71–89.

Zhang, H., Bourouts-Trites, M., 2021. The long-term washback effects of the National Matriculation English Test on college English learning in China: tertiary student per-

spectives. Stud. Educ. Eval. 68, 1–21.