Asian-Pacific Journal of Second and Foreign Language Education

RESEARCH Open Access

Challenges and opportunities for spoken English learning and instruction brought by automated speech scoring in large-scale speaking tests: a mixed-method investigation into the washback of *SpeechRater* in TOEFL iBT

Kaixuan Gong[1*]

*Correspondence:
gongkx22@mails.tsinghua.edu.cn

[1] Department of Foreign Languages and Literatures, Tsinghua University, Beijing, China

Abstract

The extensive use of automated speech scoring in large-scale speaking assessment can be revolutionary not only to test design and rating, but also to the learning and instruction of speaking based on how students and teachers perceive and react to this technology. However, its washback remained underexplored. This mixed-method study aimed to investigate the washback of TOEFL iBT Speaking's *SpeechRater* on Chinese EFL learners through questionnaire and interviews, and explore its associations with test performance and its multi-levelled influential factors. The participants received a mixture of positive and negative washback, such as their motivated individual learning through personal devices, decreasing real-life communicative practices, and increasing exam-driven behaviours. Test takers' personal understandings of automated speech scoring were found directly influential to the washback of *SpeechRater* that they experienced. Furthermore, their test scores of TOEFL iBT Speaking were positively correlated with the implicit washback of *SpeechRater* on their learning but uncorrelated with its explicit washback on their test preparation. The findings have been drawn on to propose a washback model of automated speech scoring and make suggestions to test designers, teachers and learners on how to boost its positive washback and mitigate its negative washback. This research has concluded the importance of test takers' awareness of the integrated dimensions to evaluate spoken English in real-life use. Accordingly, instructional implications are discussed on how teachers can guide the students to utilize automated speech scoring in learning and set up comprehensive learning goals for spoken English.

Keywords: Automated speech scoring, Washback, Spoken English learning, Chinese EFL learners

Introduction

Automated scoring refers to the computerised system that can evaluate the quality of test takers' performance (Williamson et al., 2006). Compared to the well-established automated writing scoring technology, the development of automated speech scoring or evaluation (ASE) was initiated only a decade ago (Evanini & Zechner, 2020). As one of the most representative and advanced ASE software, *SpeechRater* has been applied in the Speaking section of TOEFL Internet-based Test (iBT) since 2019 for its efficiency and consistency in scoring large-scale tests (ETS, 2020). Although *SpeechRater* is used in a hybrid way with human raters (ETS, 2021a), testwiseness targeting at ASE has been over-propagated in the market of TOEFL iBT test preparation and eliciting negative influence on test takers (Liu & Gu, 2013). Sceptical voices on ASE have also arisen in terms of its overdependence on linguistic features and insensitivity to content (Davis & Papageorgiou, 2021).

Given this dispute, test designers have noticed the significance of validating the application of ASE. Washback means "the effect of assessment" (Bachman & Palmer, 2010, p. 109). Washback of a scoring system, namely its influence on people involved in the test, is proposed as its crucial validity evidence (Knoch & Chapelle, 2018). Probing into the experience of test stakeholders, washback studies can enlighten the appropriate use of a scoring system (Liu, 2013) and the learning and teaching of the language skills involved in the test construct (Green, 2007). Nevertheless, the majority of previous researchers have worked with the substantial data output of machine learning to justify the reliability of ASE, while its washback remains to be evidenced.

The present study builds on this underexplored research issue with the focus on test takers. Washback has an individualised and context-specific nature, and thus research on washback should keep highly contextualised (Zhang et al., 2020). The present study focused on the Chinese test takers of TOEFL iBT Speaking in terms of how they experienced the washback of *SpeechRater*. Taking account of context specificity, this study is a small-scale one employing the undergraduates from a Chinese university.

Considering the complex nature of washback (Cheng et al., 2004), the present study applied mixed research methods to draw a complete picture of *SpeechRater*'s washback. At the quantitative stage, a questionnaire was carried out to investigate *SpeechRater*'s washback on test takers, and to examine individual differences in washback and the relationship between washback and test performance. Then, the qualitative stage adopted in-depth interviews to explore the influential factors on *SpeechRater*'s washback.

Literature review

ASE

Given the considerable workload of human raters, organisations running large-scale speaking tests have been developing and applying ASE (Lyu, 2015). It comprises three elements as shown in Fig. 1. Firstly, Speech Recognizer automatically transcribes audio files into texts. Next, Feature Computation Module can extract acoustic and linguistic features of an

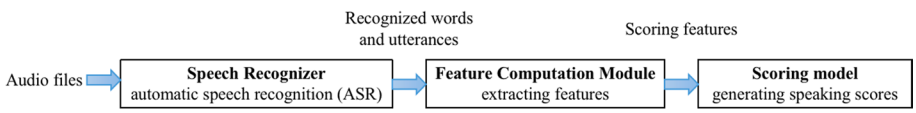

Fig. 1 The system of ASE (Zechner et al., 2009, p. 886)

input speech. This module is capable of reporting hundreds of features, among which only those with a high representation of the rating rubrics are selected and processed (Zechner et al., 2015). Finally, Scoring Model can output test scores by weighing the processed features based on its human–computer fitting model which has been set up in advance through machine learning.

ASE remains disputable among test stakeholders for its underrepresentation of test construct. In other words, ASE is commonly believed to be not as capable as human raters of evaluating speaking performance from all the dimensions including delivery, language use and content. In fact, however, the construct representation of ASE has gone through substantial evolution in recent years (Zechner & Loukina, 2020). Evanini et al.'s review (2017) notes that ASE software now can proficiently assess the surface features of pronunciation, intonation, fluency, grammar and vocabulary. The scoring of discourse coherence and content is less mature but under vigorous development. Given the all-round ASE construct, test takers still need comprehensive learning and test preparation to achieve the speaking performance that can earn high automatic scores.

Discussion on the implementation methods of ASE has focused on two issues, namely how liberal and for which task types ASE can be used (Davis & Papageorgiou, 2021). Firstly, hybrid human–machine scoring is applied in most cases (Zechner, 2020), with the degree of human involvement depending on the automated software's estimated reliability (Williamson et al., 2012). In addition, considering machine's sensitivity to language use accuracy but less to content, researchers suggest it be applied more liberally in constrained speaking tasks which elicit less spontaneous speeches (Evanini et al., 2017), such as reading-aloud and story-retelling.

When validating the application of ASE, the majority of previous studies have focused on its reliability by checking human–machine agreement (e.g., Bernstein & Cheng, 2007; Gong et al., 2009; Wang et al., 2018; Yan et al., 2009). Besides scoring precision, researchers have also examined the construct representation of ASE (e.g., Chen et al., 2018; Guan, 2019; Jin et al., 2020). Most studies justified the high reliability of their software, while researchers have brought into notice the importance of collecting validity evidence from the perspective of test takers (Williamson et al., 2012; Yang et al., 2002). Nevertheless, the studies with test takers have rarely investigated washback but only touched on their attitudes towards ASE. Test takers generally trusted that it could accurately rate the constrained tasks such as reading-aloud (Li et al., 2008). However, there were also sceptical opinions on its overreliance on phonological and grammatical features (Xi et al., 2016) and its impact on the authenticity of test settings (Fan, 2014). Given the diversified test takers' voices, it is worth going deeper to examine their experiences of ASE washback.

ASE washback in general

ASE developers and experts have predicted both its positive and negative washback. Positive washback is foreseen mainly when ASE is applied to students' self-motivated learning and practice of speaking. Bejar (2010) anticipates the increasing individual training opportunities for test takers enabled by the mobility of ASE software. Additionally, Zhang et al. (2020) expect that it can provide diagnostic feedback to test takers and facilitate their self-evaluation. Nevertheless, test preparation behaviours for high-stake

speaking assessment would usually be associated with negative washback. Xi (2010) mentions that due to ASE's sensitivity to linguistic features, test takers may pay too much attention to language use accuracy and care less about ideas and creativity in their spoken English practice. Moreover, Jin et al. (2021) show their concern about test takers' extensive training in cheating strategies to achieve high automatic scores. However, empirical evidence is still needed to demonstrate ASE washback on test takers and its influential factors.

Only two empirical washback studies so far on speaking tests have discussed the application of ASE, which implied both its positive and negative washback. The study by Yu et al. (2017) on TOEFL iBT Speaking carried out a questionnaire involving 1500 Chinese test takers on their test preparation. They found taking mock tests on TOEFL Practice Online (TPO) the most favoured test-preparation activity and the only significant predictor of test scores. With *SpeechRater* installed, TPO could simulate the scoring in TOEFL test setting and offer instant feedback. Therefore, the researchers pointed out the opportunities brought by ASE for self-evaluation during test preparation. However, Zhang's study (2019) observed test takers' exam-driven behaviours before the oral English session of high-school matriculation exam in China, and viewed them as the negative washback caused by the application of ASE. 260 students were surveyed and most of them admitted that they recited the expressions and templates which their teachers believed would contribute to high ASE scores regardless of task content. The existing findings have preliminarily uncovered the issue of ASE washback, and the present study aims to comprehensively investigate the washback of *SpeechRater* with reference to various washback dimensions.

Conceptual model and research questions

The conceptual model of the present study is based on the washback model by Bailey (1996, p. 264). Her model is developed from the washback trichotomy known as "1993, p. 2). It illustrates the mechanism of washback that tests can first adjust various test stakeholders' (*participants, processes and products*" (Hughes, *participants*) perceptions, which then influence their experience of washback (*processes*) and finally their performance (*products*). In the present model (see Fig. 2), the application of ASE is hypothesized to elicit various *processes* of test takers, who are the *participants* to be discussed. *Processes* are classified into implicit and explicit washback. As defined by Prodromou (1995, p. 14–15), implicit washback refers to test takers' daily learning of what is relevant to the skills to be tested, and explicit washback lies in targeted test preparation activities.1996; Chen, 2007; Green, 2007; Liu & Gu, 2013; Shohamy et al., 1996

Additionally, the present study aims to further explore the influential factors on the washback of ASE and conceptualise them to enrich the above model. Scholars of washback theory (Alderson & Hamp-Lyons,) have put forward four aspects of potential

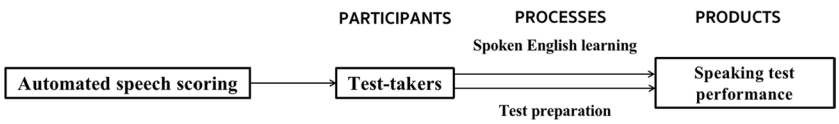

Fig. 2 Conceptual model of the present study

influential factors, including those relating to the test takers themselves, the test, the educational context and the social context. However, there is an inadequacy of empirical findings to map them out for constructing a holistic washback model.

On the basis of conceptual model and with the case of *SpeechRater*, four research questions are proposed:

RQ1 How do the test takers of TOEFL iBT Speaking view the washback of *SpeechRater* on them?

RQ2 What are the differences among the test takers with various individual characteristics in terms of the washback of *SpeechRater* on them?

RQ3 How has the washback of *SpeechRater* on the test takers influenced their test performance in TOEFL iBT Speaking?

RQ4 What are the influential factors on the washback of *SpeechRater* on the test takers?

Methods

Research context

The present study chose *SpeechRater* in TOEFL iBT Speaking for its twofold representativeness. Firstly, TOEFL iBT Speaking is a mature and popular speaking test with multiple task types and comprehensive rating dimensions. It is constituted by one independent task and three integrated tasks (ETS, 2021b, p. 25), all scored according to the rubrics of *delivery* (pacing, intonation, pronunciation and stress), *language use* (vocabulary choice, grammar use and idea cohesion) and *topic development* (completeness, organisation and appropriateness) (ETS, 2021b, p. 65–66). The rubrics of integrated tasks additionally highlight the accuracy of summarizing input materials (Huang et al., 2018). It means ASE can be more applicable to the predictable content of integrated speaking than to the open-ended answers of independent speaking, and thus the use of ASE on different tasks may elicit different washback. Secondly, *SpeechRater* has been highly developed and extensively applied. Given *SpeechRater*'s high reliability in the prior experiments by ETS (Chen et al., 2018; Xi et al., 2008), it is used to assess speaking performance from all the dimensions of TOEFL iBT Speaking rubrics, and to determine the scores together with human raters (ETS, 2021a).

China has a highly examination-oriented society (Yu & Jin, 2014) and a competitive market for TOEFL test preparation (Yu et al., 2017). Due to the common view that test results can represent academic success, Chinese TOEFL iBT test takers have been found engaging in highly purposive activities to achieve high scores. For implicit washback, for instance, Chinese test takers were much less motivated to practice casual talk at English clubs or with real people than US test takers (Ling et al., 2014). For explicit washback, the strategies of testwiseness were extremely popular and became the selling point of commercial test-preparation courses (Liu & Gu, 2013). Chinese test takers favoured the practices such as "memorizing model short essays" and inserting "certain words to lengthen their speech" (Matoush & Fu, 2012, p. 117). To mitigate negative washback and pass on the idea that progress in language proficiency is more important than high

scores, Chinese language assessment researchers have been dedicated to exploring washback mechanism (Cheng, 2008).

Research design

The present mixed-method study was based on Creswell et al.'s "sequential explanatory design" (2003, p. 180) illustrated by Fig. 3, in which the quantitative stage was followed by the qualitative stage. Questionnaire data were collected for answering RQ1, 2 and 3, and for the sampling of the following-up interviews about RQ4. As Creswell et al. (2003) recommend, the interviewees were selected from those who were characterised as the outlier individuals in questionnaire data analysis.

Sampling

As a small-scale study, this research sampled the participants from the undergraduate school of University H. It is a top comprehensive university in China that has almost 150 undergraduate majors. English language proficiency tests such as TOEFL are popular among a considerable number of the students due to their need of language certificates for studying abroad applications. According to the annual report of University H, around 15% of bachelor's degree holders every year choose to pursue further degrees overseas, and in 2020 this population was 960. Besides, every undergraduate is required to take an international exchange programme once, whether during a holiday at summer or winter schools or during one term as affiliate student. The shortlists of competitive programmes are decided based on the scores of TOEFL or IELTS. In terms of English language education, every non-English-major undergraduate has four compulsory sessions of 'college English' courses, which are carried out in IT rooms. University H, as a local test centre of TOEFL iBT, equips its IT rooms with qualified facilities. In 'college English' classes, computers are important devices for teaching and in-class writing and speaking practices.

A combination of convenience and purposive sampling strategies (Dörnyei, 2007, p. 98) was conducted. A questionnaire was spread online through an advertisement on a popular forum of University H, which was developed by the undergraduates to exchange their stories and opinions about school life. As the online forum was equally accessible for all the students, the demographic diversity of the respondents could be generally ensured. The criterion for purposive sampling as emphasized in the advertising post was

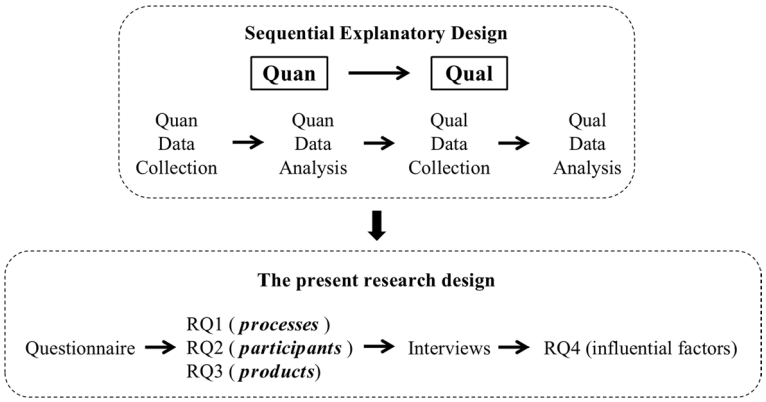

Fig. 3 Procedures of the present research design

Table 1 Descriptive statistics of the respondents' demographic information

	Category	Population
Year of study	1st year	26
	2nd year	37
	3rd year	48
	4th year and above	42
	Total	153
Gender	Female	91
	Male	55
	Prefer not to say	7
	Total	153
Major	Arts & Humanities	54
	Social Sciences	23
	Natural Sciences	16
	Engineering	32
	Information technology	11
	Agriculture, Life and Environment	10
	Medical Science	7
	Total	153

that all the respondents should once take TOEFL iBT within the last six months before they filled in the questionnaire, so that they had clear memory of their experiences. After wiping out effortless responses by conducting long-string analysis (Curran, 2016) for the scale items, 153 became valid cases among the 168 respondents. The demographic information of the participants generally demonstrated their background diversity (see Table 1).

Questionnaire

Design

The questionnaire comprised four clear and coherent sections to keep the respondents on track (see Additional file 1: Appendix A). Following Cohen et al.'s suggestion (2018, p. 493), the questionnaire started with factual questions in the Pre-section, then moved to closed questions in Section I and II, and ended with open-ended questions. To investigate the respondents' individual characteristics, the Pre-section first involved six aspects of background information with reference to the previous studies on washback variability (Allen, 2016; Bailey, 1996; Ferman, 2004; Green, 2007; Liu & Gu, 2013; Xi, 2012; Yu et al., 2017), including year of study, gender, major, principal purpose for taking TOEFL iBT, self-evaluated spoken English proficiency level, and self-evaluated familiarity with ASE. For proficiency level, as the present study recruited Chinese students, the "Self-assessment scale for oral expression" issued by National Language Commission of the People's Republic of China (2018, p. 123–124) was provided for the respondents to report their own spoken English proficiency on a scale of 1–9. Although students might not be able to exactly evaluate their actual proficiency level, this self-perception of proficiency could be closely associated with their behaviours as test washback according to socio-cognitive learning theory (Bandura,). Participants were also asked about their latest TOEFL iBT Speaking score. Given the

anonymity of questionnaire responses, they could feel comfortable to share their test results.

Section I (implicit washback) and II (explicit washback) each included 15 washback statements. These 30 statements were in the form of five-point Likert scale items, which enabled the evaluation of both washback direction and intensity (Green, 2007), as illustrated in Table 2 with one item of positive washback on learning interest and the other of negative washback on learning pressure.

Since implicit washback was less recognizable (Prodromou, 1995) and could be more difficult to recollect, the 15 implicit washback items came first when the respondents had stronger patience than they did for the subsequent 15 explicit ones. "The use of *SpeechRater*" was highlighted at the beginning of each scale item to elicit respondents' relevant memory and perceptions. The indicators of ASE washback were decided on by referring to previous washback studies (Alderson & Wall, 1993; Bailey, 1996; Barnes, 2016; Bejar, 2010; Chen, 2007; Ferman, 2004; Jin, 2000; Yu et al., 2017; Zhang, 2019) and ASE studies (Jin et al., 2020; Xi, 2010; Xi et al., 2016). In this way, implicit washback was represented by learning content, time, materials, activities, interest, pressure and targets, and explicit washback by test-preparation behaviours and effects. Specifically, for the scale items of learning content, the involved language skills all came from the rating rubrics of TOEFL iBT Speaking, for examining whether the use of *SpeechRater* encouraged test takers' learning of what was expected by the TOEFL construct.

Finally, at the end of questionnaire there were two open-ended questions, one for the respondents to share the opinions they would like to emphasise besides the fixed items, and the other for them to provide their contact information if they would like to be potential interviewees.

The questionnaire went through careful piloting and revision before it was disseminated. Cronbach's reliability alpha analysis was applied to verify that the scale items functioned as intended. It indicated that the overall 30 scale items ($\alpha = 0.843$), the 15 implicit washback items in Section I ($\alpha = 0.855$) and the 15 explicit washback items in Section II ($\alpha = 0.708$) all featured high internal reliability.

Table 2 Illustration of what each rank in a five-point Likert scale represented

	1 (strongly disagree)	2 (disagree)	3 (neutral)	4 (agree)	5 (strongly agree)
The use of SpeechRater in TOEFL iBT Speaking test has promoted my interest in spoken English learning	Intense negative washback	Negative washback	Neutral	Positive washback	Intense positive washback
The use of SpeechRater in TOEFL iBT Speaking test has brought me pressure in spoken English learning	Intense positive washback	Positive washback	Neutral	Negative washback	Intense negative washback

Questionnaire data analysis

To answer RQ1 (washback on test takers), the descriptive statistics for each of the 30 washback items were generated with SPSS 27.0, including the mean score and standard deviation of the respondents' answers. For the open-ended questions on the respondents' opinions, keyword analysis was conducted by extracting the five most frequently mentioned concordances with the Word List and Concordance tools of AntConc 3.5.9.

RQ2 (washback variability) would be answered by estimating the correlations between the respondents' answers to the 30 washback items and their six aspects of individual characteristics. ANOVA tests (for evaluating the differences among multiple groups) were applied to the four categorical variables (gender, year of study, major, principal test purpose), and regression analysis (for evaluating the relationships between multiple numerical variables) to the variables of spoken English proficiency level and familiarity with ASE.

For RQ3 (the relationship between washback and test performance), a partial correlation analysis was conducted first. To be specific, with the respondents' spoken English proficiency level as the covariate, the correlations between their answers to the 30 washback items and their TOEFL iBT Speaking scores were examined. After the washback items significantly correlated with the test scores were found, a multiple linear regression analysis was applied then to detect the predictors of test scores with spoken English proficiency level as the control variable.

Interviews

After the answers of the scale items on negative washback were reversed, the sum of 30 washback items' scores of each questionnaire respondent was calculated. Two individuals with the highest scores and two with the lowest were selected as the four interviewees (see Table 3). In this sense, Case 1 and 2 had experienced the most positive washback

Table 3 Background information of the four interviewees

	Washback	Gender	Year of study	Major	Spoken English proficiency (Band 1–9)	Principal test purpose	Familiarity with ASE
Case 1-Ge	Positive	Male	1st year	Information Technology (Computer Science)	Band 4	For international exchange programmes	Very familiar
Case 2-Yan	Positive	Female	4th year	Social Science (International Studies)	Band 6	For presenting test result to potential employers	Neutral
Case 3-Ling	Negative	Female	2nd year	Arts & Humanities (English Language & Literature)	Band 8	For evaluating spoken English proficiency	Very unfamiliar
Case 4- Kun	Negative	Male	4th year	Natural Science (Biology)	Band 5	For studying abroad for further degrees	Unfamiliar

of *SpeechRater*, and Case 3 and 4 the most negative. After knowing about the present study and their rights as interviewees, they all agreed to be recruited.

The interview protocols (see Additional file 1: Appendix B) were constituted of two themes: bottom-up questions and top-down questions. Specifically, each interviewee was invited to first describe and reflect on their own spoken English learning and test preparation given the application of *SpeechRater*, and then identify the influential factors on its washback. To facilitate interviewees' thinking and to bridge the social distance between them and the interviewer, the interviews were carried out in Chinese, their shared native language. Each interview took around one hour online and was recorded. After the recordings were transcribed verbatim in Chinese, the transcripts were given back to the interviewees to confirm. Only the excerpts included in this paper were translated into English word by word.

RQ4 (the influential factors on washback) would be addressed by analysing the interviews. To directly refer to previous washback studies, coding of the interview data was conducted in English and through *qualitative content analysis* (Mayring, 2004). It emphasised the exploration of key information through setting up the categories for coding on the basis of revising pre-determined categorisation during data analysis (p. 269). The coding process involved four steps. Firstly, the sentences that implied the influence of certain factors on *SpeechRater*'s washback were highlighted, with cause-effect words as the key identifiers for the bottom-up questions' responses. Secondly, the highlighted excerpts were paraphrased into straightforward cause-effect sentences. Thirdly, the paraphrased sentences were generalised into impersonalised statements. Finally, the statements were labelled and classified into one category of influential factors. There had been four existing categories of influential factors on washback, including Test takers, Test, Educational context and Social context. Considering the present study's focus on the washback of the scoring system rather than the whole test, a fifth category was newly created for the influential factors relating to *SpeechRater* per se. To further ensure reliability, the coding results were reviewed by an experienced English teacher at University H from China.

Results and discussion

Test takers' views on the washback of *SpeechRater* in TOEFL iBT Speaking

Implicit washback

In terms of learning content, the questionnaire respondents were found modestly motivated to learn all six language skills. Cohesive devices (M = 3.86, SD = 0.884) and pronunciation (M = 3.83, SD = 0.785) won the largest proportion of motivated learners, and closely behind them were vocabulary (M = 3.74, SD = 0.849) and intonation (M = 3.71, SD = 0.826). The motivated learning of grammar (M = 3.65, SD = 0.870) was relatively mild, and the respondents were least motivated to develop pragmatic competence (M = 3.52, SD = 0.836). The test takers' unbalanced learning of different language skills went in line with Xi et al.'s findings (2016) about students' perception of *SpeechRater*. It indicated that test takers' awareness and understanding of the existence of ASE could strongly mediate their received washback. In their minds, the results of ASE depended more on their delivery and language use than on their idea and appropriateness. Given this unbalanced representation of different aspects of the speaking test's construct,

automatically-scored test takers could pay less attention to high-order English speaking abilities in their learning.

Furthermore, both the results of mildly increasing learning pressure (M = 3.34, SD = 0.994) and decreasing interest (M = 2.82, SD = 1.007) indicated the negative washback of *SpeechRater*. However, slight positive washback was implied by the answers to the two items on its effect of locating deficiencies (M = 3.44, SD = 0.986) and setting goals of spoken English learning (M = 3.45, SD = 0.966). Also, the time (M = 3.45, SD = 0.859) and materials (M = 3.46, SD = 0.993) for spoken English learning increased moderately. Finally, among the off-class learning activities, the increasing use of the smartphone applications relating to spoken English learning and testing (M = 3.65, SD = 0.997) was the most evident. By contrast, *SpeechRater* did not motivate the respondents to participate in face-to-face English speaking. Specifically, the activities that involved casual talk (M = 2.88, SD = 0.975) received less attention than those in formal academic settings (M = 3.07, SD = 1.017).

SpeechRater elicited a blend of positive and negative washback on learning activities. Given automated scoring software's mobility (Bejar, 2010) and usefulness in providing diagnostic feedback (Zhang et al., 2020), it could enlighten test takers to practice spoken English independently through personal devices, facilitating their realisation of learning initiative. However, many test takers were demotivated to join in-person English communications, which in their minds could not directly contribute to their performance in the automatically-scored tests where no human listeners were around judging them instantly. This meant ASE interrupted the seamless move from practical activities to exam tasks (Fan, 2014). Its application affected the authenticity of speaking tests, a crucial prerequisite of positive washback (Messick, 1996).

Explicit washback

The explicit washback of *SpeechRater* was overall more intense than its implicit washback as the statistics illustrated. This corresponded to Prodromou's statement (1995, p. 14–15) that explicit washback could be more purpose-driven and thus more recognizable. Among the changes in the respondents' test-preparation behaviours given the use of *SpeechRater*, locating the grammatical mistakes of their own speech during self-evaluation ranked the highest in popularity (M = 4.11, SD = 0.757). This implied that they perceived *SpeechRater* as highly sensitive to grammatical accuracy. Automated software essentially rated the textual version of speech, however, spoken discourse could hardly be as grammatically-structured as written discourse (Xi, 2010). Consequently, the strictness of ASE in language use accuracy could yield a double-edged effect. It motivated test takers to improve their grammar use, but distracted them from the emphasis on comprehensibility by spoken English in authentic situations.

The other strongly favoured test-preparation behaviours (M > 4) were all closely test-related, including doing mock tests (M = 4.10, SD = 0.690), studying TOEFL iBT Speaking rubrics (M = 4.04, SD = 0.760), and practicing talking to oneself on the topics that frequently appeared in past tests (M = 4.10, SD = 0.779). Test takers' favour of exam-driven behaviours was explained in the previous washback studies with their priority given to test results rather than making progress (Yu et al., 2017; Zhang, 2019). This negative washback was exacerbated in the present study, as the respondents perceived their

test preparation as more effective for gaining a higher score from *SpeechRater* (M = 3.99, SD = 0.730) than for improving spoken English proficiency (M = 3.61, SD = 0.845). This perception consequently intensified explicit washback and restricted their learning within fixed and monotonous task types when preparing to be automatically scored.

Finally, the five most frequently-mentioned concordances in the open-ended question answers included: reliance on exam-driven activities; demotivation to practice casual talk; importance of improving grammatical accuracy; demotivation to practice in authentic English-speaking settings; usefulness of smartphone applications that can evaluate their speaking performance. These all went consistently with the prominent statistical results mentioned above.

Individual differences in the washback of *SpeechRater*

The results of ANOVA tests demonstrated the washback variability as follows. Firstly, there was no significant difference between the two genders. On the dimension of years of study, different degrees of preference for short-term test preparation to long-term practice were found ($F(3, 73) = 3.770$, $p < 0.05$, partial $\eta^2 = 0.134$). LSD Post Hoc test further indicated that the undergraduates in 2nd year and 4th year and above preferred short-term test preparation more than 1st year ones did ($p < 0.05$). Yu et al. (2017) also observed more intense negative explicit washback of TOEFL iBT Speaking among the older group, who attached greater importance to test results while the younger group expected to make genuine progress. Likewise, in the present context of University H, most 2nd year students took TOEFL to apply for exchange programmes and 4th year students for studying abroad. Their urgent need for high scores to fulfil the requirement made their test preparation more purpose-driven. In addition, different principal test purposes were found bringing the variation of learning pressure ($F(4, 73) = 3.166$, $p < 0.05$, partial $\eta^2 = 0.148$). Those who took TOEFL to apply for overseas study and international exchange programmes endured more pressure than those for evaluating their spoken English proficiency ($p < 0.05$). The above findings implied that test takers' feeling of test importance could make a difference to the washback of *SpeechRater* they received.

Besides, the respondents from various majors showed significant differences in terms of reciting templates ($F(6, 73) = 3.496$, $p < 0.05$, partial $\eta^2 = 0.126$). LSD Post Hoc test results showed that both Natural Sciences and Social Sciences students agreed more than information technology (IT) students that templates were helpful ($p < 0.05$). IT students' doubt in templates was further explained by Ge (Case 1), the IT-major interviewee, as their awareness that automated scoring software was trained with big data and might give low scores to the identified templates. It could be inferred that washback was closely relevant to test takers' knowledge and understanding of ASE (Xi, 2012).

Finally, as the results of regression analysis illustrated (see Table 4), the respondents with lower spoken English proficiency levels focused more on the learning of surface features including pronunciation, intonation and grammar. Lower-level respondents also indicated their higher tendency to recite templates and their more impeded normal learning activities due to test preparation. It implied that they could perceive additional test demand given the use of *SpeechRater* and consequently exacerbated negative explicit washback.

Table 4 Multiple Linear Regression analysis results of significant washback variability

Statement (dependent variable)	Spoken English proficiency level (independent variable)	R^2
(1) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *learn English pronunciation*	−0.082* (−2.288)	0.040
(2) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *learn English intonation*	−0.145*** (−3.975)	0.100
(4) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *learn English grammar*	−0.102** (−2.727)	0.068
(17) As TOEFL iBT Speaking test is automatically-scored, before the test it has been helpful for me to ***r****ecite some templates*	−0.099* (−2.515)	0.073
(24) As TOEFL iBT Speaking test is automatically-scored, before the test it has been helpful for me to enrol in *commercial courses* on the test preparation for TOEFL iBT Speaking	−0.110** (−2.728)	0.050
(28) As TOEFL iBT Speaking test is automatically-scored, through preparation for TOEFL iBT Speaking test, I *improved my oral English proficiency*	−0.136*** (−3.672)	0.094
(29) As TOEFL iBT Speaking test is automatically-scored, through preparation for TOEFL iBT Speaking test, I *was impeded in terms of my normal spoken English learning activities*	0.109* (2.426)	0.046
Statement (dependent variable)	**Familiarity with ASE (independent variable)**	**R^2**
(5) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *learn English cohesive devices*	0.151* (2.062)	0.045
(7) The use of *SpeechRater* in TOEFL iBT Speaking test has made me *spend more time on spoken English learning and practice*	0.202** (2.741)	0.059
(8) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to ***l****earn with audio/video English materials*	0.230** (2.823)	0.070
(9) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *take part in off-class spoken English communication activities*	0.273** (3.463)	0.094
(10) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *speak more English in academic settings*	0.220** (2.679)	0.098
(13) The use of *SpeechRater* in TOEFL iBT Speaking test has *promoted my interest* in spoken English learning	0.171* (2.031)	0.033
(14) The use of *SpeechRater* in TOEFL iBT Speaking test has helped me *locate my deficiencies* in spoken English learning	0.296*** (3.711)	0.097
(15) The use of *SpeechRater* in TOEFL iBT Speaking test has *set clear spoken English learning goals* for me	0.241** (3.055)	0.078

*p < 0.05, **p < 0.01, ***p < 0.001; *t* statistics in parentheses

Notably, the individual washback differences caused by the respondents' various familiarity with ASE all lay in implicit washback (see Table 4). The more familiar they were with ASE, the more possible they were found to experience positive washback on spoken English learning, including their increasing learning time, materials, activities and interest. Allen's study (2016) on test washback once indicated a positive correlation between familiarity with the test and positive washback. When it came to the washback of ASE, the importance of test takers' relevant knowledge was also highlighted by expert test designers (Evanini & Zechner, 2020) and corroborated by the present study. The more test takers understand the technology, construct and application of ASE, the more likely they could keep a robust spoken English learning style rather than being impeded by blind and excessive test preparation.

The influence of the washback of *SpeechRater* on test performance in TOEFL iBT Speaking

With the respondents' spoken English proficiency level as the covariate, five washback items were found significantly correlated with TOEFL iBT Speaking scores in the partial correlation analysis (see Table 5). Four of the five items belonged to implicit washback and were all positively correlated with test scores, which meant active learning had effectively improved test performance. The only explicit washback item, reciting templates, was observed as negatively correlated with test scores. Similarly, the study by Yu et al. (2017) on test preparation of TOEFL iBT Speaking concluded a weak relationship between most explicit washback and test performance. The present study, by involving the investigation of both explicit and implicit washback, further justified that learning was still more important than test preparation to determine automatic scores.

A multiple linear regression analysis was then conducted with these five items as the independent variables, spoken English proficiency level as the control variable, and test scores as the dependent variable (see Table 6). 43.1% of test scores ($F(6, 146) = 20.209$, $p < 0.001$) could be explained by this model, which suggested that it worked well. With self-assessed spoken English proficiency level being the most powerful predictor

Table 5 The washback items that had statistically significant correlations with TOEFL iBT Speaking scores

Statement	*r* (Pearson)	*Sig.*
(5) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *learn English cohesive devices*	0.189	0.020
(9) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *take part in off-class spoken English communication activities*	0.195	0.016
(10) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *speak more English in academic settings*	0.179	0.027
(11) The use of *SpeechRater* in TOEFL iBT Speaking test has motivated me to *use the smartphone applications relating to spoken English learning and testing*	0.293	0.000
(17) As TOEFL iBT Speaking test is automatically-scored, before the test it has been helpful for me to *recite some templates*	−0.193	0.017

Table 6 Multiple Linear Regression of TOEFL iBT Speaking test scores and washback items

	Unstandardized coefficients		Standardized coefficients	*t*	*Sig.*
	B	*SE*	Beta		
(Constant)	15.163	1.288		11.771	0.000
(5) The use of *SpeechRater* in TOEFL iBT speaking test has motivated me to *learn English cohesive devices*	0.362	0.200	0.118	1.806	0.073
(9) The use of *SpeechRater* in TOEFL iBT speaking test has motivated me to *take part in off-class spoken English communication activities*	0.306	0.226	0.110	1.353	0.178
(10) The use of *SpeechRater* in TOEFL iBT speaking test has motivated me to *speak more English in academic settings*	−0.093	0.228	−0.035	−0.408	0.684
(11) The use of *SpeechRater* in TOEFL iBT speaking test has motivated me to *use the smartphone applications relating to spoken English learning and testing*	0.548**	0.187	0.202	2.934	0.004
(17) As TOEFL iBT speaking test is automatically-scored, before the test it has been helpful for me to *recite some templates*	−0.354	0.195	−0.115	−1.816	0.071
Self-assessed spoken English proficiency level	0.887***	0.090	0.636	9.803	0.000

($p < 0.001$), there was one washback item that significantly predicted test scores: the respondents who were more motivated given the use of *SpeechRater* to practice spoken English with smartphone applications had performed better in TOEFL iBT Speaking ($p < 0.05$). In Yu et al.'s regression analysis (2017), they reported practicing with TPO as the only predictor of test performance. TPO and smartphone applications were similar in terms of providing instant automatic scores and diagnostic feedback. In this way, test takers could familiarise themselves with the construct, reliability and sensitivity of ASE, and experience the feeling of speaking to a nonhuman rater. Despite the similarity, their essential difference was that test takers' use of TPO was test-driven, while using smartphone applications was learning-driven. TPO had fixed task types and test formats, while those in smartphone applications were much more diversified. In this sense, ASE might encourage and further enable independent learning and self-evaluation, and the test takers taking advantage of it could both get high scores and make progress in spoken English.

Multi-levelled influential factors on the washback of *SpeechRater*

After sorting all the respondents' scale item answers, the four extreme cases were invited to the subsequent interviews: Ge (Case 1) and Yan (Case 2) had the most positive views of washback, while Ling (Case 3) and Kun (Case 4) the most negative. The findings with them were categorised into the following subsections according to the five levels of influential factors on the washback of *SpeechRater*.

Test takers level: the personal factors

Ge knew well about the technology of automated scoring as majoring in Computer Science and stayed positive about its washback. Since *'SpeechRater can follow a strict framework and take all the rating dimensions into consideration'*, he noted his comprehensive learning of spoken English by practicing delivery, vocabulary, grammar and idea organisation. He also appreciated *SpeechRater* for its objectivity, which resolved his concern about human raters' bias on his non-nativelike accent and alleviated his anxiety during both test-taking and daily learning. In terms of test preparation, Ge expressed his strong preference for long-term spoken English practice to exam-driven behaviours such as reciting templates. In his words, he had sufficient time and limited pressure as a 1st year undergraduate, and he believed that *SpeechRater* could recognise the use of templates and give low marks.

Yan shared Ge's positive views of the reliability and the washback of *SpeechRater*. On her spoken English learning, she mentioned her abundant opportunities to practice oral communication as an intern at the office of international affairs at University H. As *'authentic settings emphasised comprehensibility'*, once coming to the automatically-scored test, she began to pay attention to the accuracy of language use, which in her eyes was highly beneficial. Majoring in international studies, she dedicatedly learned and practiced spoken English for her future career as a diplomat. Therefore, as proficiency meant more to her than a high score, she did not prepare tricks or templates targeted at automated scoring, which she felt could disturb her natural performance.

Ling held a highly negative attitude towards *SpeechRater* and experienced its negative washback. Among the four interviewees, she evaluated herself as proficient in spoken

English the best while knowing ASE the worst according to her questionnaire response. Her impression of *SpeechRater* was being *'too strict on surface features and too insensitive to content'*. As she admitted, this became the reason why she prepared *'beautiful templates'* despite her dislike of fixed answers. In Ling's reflection, as a discreet first-time test taker to be automatically scored, she relied too much on preparing test-taking strategies but her performance was still devalued by *SpeechRater*. Moreover, she could tell that her practice of templates affected her argumentation and creativity when making improvised speeches in her English-major classes and as a member of the English Speech Community.

Kun also pointed out the strongly negative washback of *SpeechRater* as Ling did, while he had a lower proficiency level and a more urgent test purpose of getting the TOEFL certificate for postgraduate applications in the current term. As he acknowledged, *'I knew my test preparation had limited benefit to my spoken English learning, but that was the most efficient way for me to get a high score'*. Since *'machine was essentially less intelligent than human'*, he spent much time preparing test-taking strategies and complained that the use of *SpeechRater* had distracted him from authentic spoken English practice. He became reluctant to participate in face-to-face spoken English activities, which were less effective to improve his test performance.

To summarize, there were three themes of personal factors, namely the interviewees' background information, learning style and technology-related knowledge. Firstly, test takers' background information (year of study, major, spoken English proficiency, test-taking purpose) and how these factors influenced *SpeechRater*'s washback were generally in alignment with the analysis for RQ2. Additionally, nativelikeness was a novel influential factor explored by the interview. Compared with human raters who would naturally keep a social distance from the test takers, ASE under big data training was perceived as more impartially receptive to different accents and caused less learning pressure. The second theme was about test takers' learning-related factors, including their intention of learning spoken English and willingness of seizing practice opportunities. There was a sharp contrast between the story of Yan and Kun and how they viewed the washback of *SpeechRater*, which indicated the importance of steadfast learning goals and active participation in authentic settings. These could help test takers set up a clear mind of what excellent speaking performance should be like and sensibly plan their learning rather than pursuing testwiseness. Lastly, the most evident difference between the pair receiving positive washback (Ge and Yan) and the pair receiving negative washback (Ling and Kun) lay in their divergent knowledge levels of ASE. The test takers who were more familiar with the technology had felt more determined in setting goals for their learning and test preparation.

Scoring level: the factors in SpeechRater

While the advantages of *SpeechRater* were highlighted by those who received positive washback, those who experienced negative washback expressed their dissatisfaction about its defects. Yan appreciated the mobility of ASE and was then inspired to practice spoken English on a smartphone application that could instantly generate analytic scores and diagnostic feedback on her performance. In her view, the use of *SpeechRater* facilitated her independent learning and self-inspection. In contrast, Ling and Kun both

noticed *SpeechRater*'s reliance on automatic speech recognition (ASR), which would faithfully transcribe every phoneme and intolerantly rate hesitation and self-revision. It could lead to learners' *'higher pressure'* (Ling), and Kun also mentioned his decreasing confidence in spoken English learning, as he had become accustomed to preparing a speech by *'writing down and reciting everything to ensure the accuracy'*.

To explore the interviewees' opinions of how the implementation method of *SpeechRater* could influence its washback, they were asked about any proposals to improve its current application to bring about positive effects for their learning. Ling and Kun both suggested increasing the human rating proportion in human–machine hybrid. Kun said *'the point is that the test takers and their teachers should be aware that human raters would also play a significant role to decide on the scores'*. As Williamson et al. (2012) explained, how much the machine was involved in the scoring system could determine how much test stakeholders' perceptions of automated scoring would influence its washback on them. In terms of the task types to apply *SpeechRater* to, Ge and Ling raised that the automatic scores should count less for the independent task than for the integrated tasks. They viewed *SpeechRater* as more applicable to integrated speaking, the summary of pre-determined information, than independent speaking which valued test takers' own ideas. The consistency between the scoring system's and the speaking test's constructs is a significant determiner of washback (Green, 2007). Considering the sensitivity of ASE to language use accuracy but less to content, test takers found it more suitable for constrained speaking task types, which was also indicated in Li et al.'s study (2008) on a reading-aloud task.

Test level: the factors in TOEFL iBT Speaking

The interviewees noticed various features of the test format and context that could aggravate the negative washback of *SpeechRater*. Given the time limit of each task, Kun worried that he could be too anxious to keep his language use perfect for *SpeechRater*. Consequently, he tried to memorise templates word by word beforehand to avoid making mistakes. The strict test format controlled with computers in TOEFL iBT Speaking had intensified test takers' anxiety about being automatically scored, which consequently increased their test preparation activities. Besides, Yan as an experienced English speaker in authentic situations found the TOEFL iBT test setting unnatural and oppressive. In one test room, test takers spoke simultaneously to the machine and could hear the others' voices. For face-to-face speaking tests, test authenticity is largely associated with task interactivity (Filipi, 2015). In this sense, the further limited test authenticity in ASE settings could further lead to test takers' demotivation to take real-life spoken English practices given the use of *SpeechRater*.

Educational context level: the factors in the context within University H

Based on their own knowledge and experiences, the four interviewees also demonstrated different focuses on the contextual factors at University H. Ge, who took TOEFL to apply for international exchange, stressed the institutional policy that TOEFL scores would be referred to for selecting the candidates of competitive exchange programmes. Consequently, the applicants striving for high automatic scores became more likely to conduct excessive exam-driven behaviours. However, Yan pointed out that University

H, as one of the local test centres of TOEFL iBT, provided the students with advanced equipment for computer-based learning and practice. These "resources to meet the test demands" (Green, 2007, p. 24) could bring positive washback to the test takers.

Furthermore, the students were found referring to teaching activities to understand how to evaluate spoken English, which affected how they accepted *SpeechRater* and reacted to its application. Ling, an English major student, mentioned that her major courses were *'seldom carried out in IT rooms'* and her teachers always *'highlighted critical and creative thinking in her spoken English performance'*, which caused her distrust of ASE. In contrast, Kun was a Biology student who only took 'college English' courses twice a week and had limited chances of in-class speaking practice due to the large class size. Therefore, he felt bewildered with how to evaluate speaking performance and what to improve for taking the test, and could only resort to test-taking strategies when faced with *SpeechRater*. This said, teachers should shoulder the responsibility of forming the students' right conception of spoken English evaluation and setting their comprehensive learning goals.

Although Shohamy et al. (1996) recognised school curriculum as a significant influential factor for washback, it was not observed in the present study since TOEFL iBT Speaking is a proficiency test with little relation with in-class teaching content.

Social context level: the factors in Chinese society

In comparison with Educational context level, the factors about Social context that influenced *SpeechRater*'s washback were in less proximity with the interviewees as socially-inexperienced university students. Under China's context as a thriving market for TOEFL iBT test preparation (Yu et al., 2017) and an examination-oriented society (Yu & Jin, 2014), the test takers recognised two social factors. Firstly, Kun repeatedly criticised TOEFL iBT Speaking preparation courses in the market:

> *'Teachers running those courses always highlighted that we would be assessed by machine. They spent most time introducing test-taking techniques, such as piling up advanced expressions and practicing speech rate. They boasted about these techniques as the selling points of their courses.'*

Consequently, the commercial courses brought him a negative attitude towards ASE and more confusion about how to improve spoken English proficiency. The overemphasised testwiseness by the market affected test takers' learning autonomy, which was named by Liu and Gu (2013) as the most evident influential social factor on washback. Secondly, as Yan was busy with job applications, she noticed that TOEFL iBT score reports were widely accepted by employers as a sound certificate of English proficiency level. Therefore, the test takers might tend to pursue high marks rather than genuine improvement in spoken English. In this sense, the competitive job market issued the social version of 'institutional policy' that led to excessive exam-driven behaviours targeted at ASE.

By summarising the above-mentioned findings and discussion, a washback model of ASE (see Fig. 4) was constructed. It integrated the multi-levelled influential factors with the conceptual washback mechanism of '*participants-processes-products*' in Fig. 2.

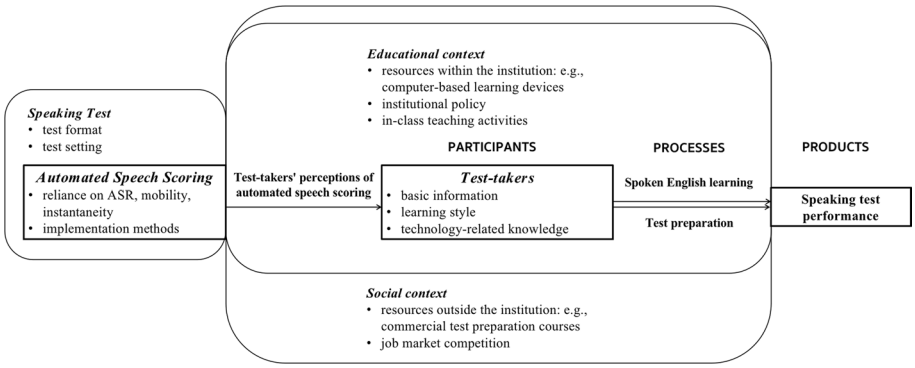

Fig. 4 A washback model of ASE

Conclusion

The present study on ASE for the first time collected its validity evidence from the perspective of test takers and investigated its washback. With the case of *SpeechRater* in TOEFL iBT Speaking, and adhering to the mechanism of *'participants-processes-products'*, a mixed-method approach was carried out to draw a holistic picture of the washback of ASE.

On the whole, the present study has observed a mixture of positive and negative washback of *SpeechRater*. Its implicit washback included the test takers' unbalanced learning focus allocated to different spoken English skills, and their motivation to engage in individual learning but demotivation in real-life communicative activities. The explicit washback was embodied by their increasing exam-driven behaviours to raise automatic scores. It was indicated that the washback of ASE experienced by the test takers was closely relevant to their personal perceptions of this technology's sensitivity to various rating dimensions. Washback variability was then found among the test takers in different majors, years of study, test purposes and proficiency levels. Notably, the present study also proved that the more test takers were familiar with the technology and application of ASE, the more likely they could receive its positive implicit washback, namely well-planned spoken English learning. Furthermore, on the relationship between washback and test scores, the correlation analysis showed that despite the use of ASE, test performance was still closely related to test takers' learning-driven practices rather than test-driven preparation behaviours. Furthermore, the only significant predictor of test performance among washback items was practicing on the smartphone applications for spoken English learning and testing, which represented the learning initiative of automatically-scored test takers. Finally, the present study explored and interpreted the influential factors on the washback of ASE, which were categorised into the five levels of Test takers, Scoring system, Test, Educational context and Social context. They imposed a synergetic effect with the factors on Test takers level mediating the other levels' influence on washback.

The present study has shed light on the washback mechanism of ASE and proposed a theoretical model, which can facilitate the research design of future washback studies on automated scoring systems. On the practical level, the findings of the present study can enlighten test designers, teachers and learners on how to boost the positive washback and mitigate the negative washback. Test designers can diversify the

implementation methods of ASE for different task types in one test, e.g., more human involvement for less constrained tasks. They should also try their best to provide clear test specifications about how ASE is implemented in each task, and collect preliminary feedback from test takers and instructors to avoid potential misunderstandings. In the same vein, the first suggestion for teachers and learners is to get familiarised with the rubrics and application of ASE to sensibly plan learning and test preparation. Secondly, teachers can engage students in spoken English activities in real-life situations for their clear understanding of all-round dimensions to evaluate spoken English and their comprehensive learning goals. Furthermore, peer assessment activities can be organised for students to perceive the process of authentic rating. Finally, teachers can also encourage and instruct students' use of ASE on mobile devices for independent practice and self-evaluation of spoken English.

Admittedly, there are limitations of the present study for future research to make up. Firstly, as a cross-sectional study, this research has included no comparison to pre-operational use of *SpeechRater* or to other human-rated speaking tests, and future contrastive studies are still needed. Secondly, the findings about *SpeechRater* are not necessarily transferrable to all ASE software, and neither is TOEFL iBT Speaking to all speaking tests. *SpeechRater* works in a human–machine hybrid way for the task types of independent and integrated speaking. Future studies can work on the tests with different implementation methods of ASE and with diversified speaking tasks, such as reading-aloud and pair work. Thirdly, due to the small-scale and highly-contextualised nature of the present study, its sampling cannot represent the whole population. Future researchers can investigate the population from the contexts other than universities, such as non-elite colleges, high schools and job settings.

Abbreviations

ASE	Automated speech evaluation
ASR	Automatic speech recognition
C/IE responses	Careless or insufficient effort responses
EFL	English as foreign language
IT	Information technology
RQ	Research question
TOEFL	Test of English as a Foreign Language
TOEFL iBT	TOEFL internet-based test
TPO	TOEFL practice online

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40862-023-00197-2.

Additional file 1. Appendix A (Questionnaire) and Appendix B (Interview protocols).

Acknowledgements
The author would like to thank Dr. Denise Chappell at University of Cambridge for her careful supervision.

Author contributions
KG designed the research and wrote the manuscript after data collection and analysis. The author read and approved the final manuscript.

Funding
This work was supported by China Scholarship Council and Cambridge Trust.

Availability of data and materials
The datasets generated during and analysed during the current study are not publicly available as noted in the privacy agreement with the participants, but are available from the corresponding author on reasonable request.

Declarations

Ethical approval and consent to participatelarations
With reference to the Ethical Guidelines for Educational Research by British Educational Research Association (2018), this study has taken into account the participants' rights, privacy and convenience. All the participants involved in the present study were adults, who were able to make independent judgements.

Competing Interests
Not applicable.

Received: 18 November 2022 Accepted: 19 May 2023
Published online: 15 August 2023

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297.
Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.
Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*(1), 1–20.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing and using language assessments in the real world*. Oxford University Press.
Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279.
Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*(9), 1175–1184.
Barnes, M. (2016). The washback of the TOEFL iBT in Vietnam. *Australian Journal of Teacher Education, 41*(7), 158–174.
Bejar, I. I. (2010). Can speech technology improve assessment and learning? New capabilities may facilitate assessment innovations. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.365.2727&rep=rep1&type=pdf
Bernstein, J., & Cheng, J. (2007). Logic and validation of fully automatic spoken English test. In M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 174–194). Routledge.
British Educational Research Association. (2018). *Ethical guidelines for educational research (4th ed.)*. Retrieved from https://www.bera.ac.uk/publication/ethicalguidelines-for-educational-research-2018.
Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., & Mundkowsky, R. (2018). Automated scoring of nonnative speech using the *SpeechRater*^{SM} v. 5.0 engine. *ETS Research Report Series, 2018*(1), 1–31.
Chen, X. (2007). On washback in language testing. *Journal of PLA University of Foreign Languages, 30*(3), 40–44.
Cheng, L. (2008). The key to success: English language testing in China. *LanguageTesting, 25*(1), 15–37.
Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Lawrence Erlbaum Associates Inc.
Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.
Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240). Sage.
Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.
Davis, L., & Papageorgiou, S. (2021). Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assessment in Education: Principles, Policy & Practice, 28*(4), 437–455.
Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.
ETS. (2020). *TOEFL® Research Insight Series*, Volume 2: TOEFL Research. Retrieved from https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v2.pdf
ETS. (2021b). *The TOEFL iBT® Test Prep Planner*. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_student_test_prep_planner.pdf
ETS. (2021a). The *SpeechRater®* Service. Retrieved fromhttps://www.ets.org/accelerate/ai-portfolio/speechrater
Evanini, K., Hauck, M. C., & Hakuta, K. (2017). Approaches to automated scoring of speaking for K–12 English language proficiency assessments. *ETS Research Report Series, 2017*(1), 1–11.
Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 3–20). Routledge.
Fan, J. (2014). Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach. *Language Testing in Asia, 4*(1), 1–17.
Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 191–210). Lawrence Erlbaum Associates.
Filipi, A. (2015). Authentic interaction and examiner accommodation in the IELTS speaking test: A discussion. *Papers in Language Testing and Assessment, 4*(2), 1–17.
Gong, L., Liang, W., & Ding, Y. (2009). Feasibility study and practice of machine scoring of repetition questions in large-scaled English oral test. *Computer-Assisted Foreign Language Education, 126*, 10–15.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge University Press.
Guan, Y. (2019). *A validation study on automated scoring of English listening and speaking test in college entrance examination of Guangdong province*. [MA Thesis]. Guangdong University of Foreign Language Studies.
Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27–49.
Hughes, A. (1993). *Backwash and TOEFL 2000*. [Unpublished manuscript]. University of Reading.
Jin, Y. (2000). Washback effect of CET-SET on teaching and learning. *Foreign Language World, 2000*(4), 56–61.
Jin, Y., Wang, W., & Yang, H. (2021). Technological applications in language assessment: A case study of College English Test. *Foreign Language Testing and Teaching, 2021*(1), 1–27.
Jin, Y., Wang, W., Zhang, X., & Zhao, Y. (2020). A preliminary investigation of the scoring validity of the CET-SET automated scoring system. *China Examinations, 339*, 25–33.
Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499.
Li, M., Yang, X., Feng, G., Wu, M., Chen, J., & Hu, G. (2008). Feasibility study and practice of machine scoring of reading aloud item in large-scale college English oral tests. *Foreign Language World, 2008*(4), 88–95.
Ling, G., Powers, D. E., & Adler, R. M. (2014). Do TOEFL iBT® scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument. *ETS Research Report Series, 2014*(1), 1–16.
Liu, J. (2013). Modern educational technology and language testing. *Computer-Assisted Foreign Language Education, 152*, 46–51.
Liu, O. L. (2014). Investigating the relationship between test preparation and TOEFL iBT® performance. *ETS Research Report Series, 2014*(2), 1–13.
Liu, X., & Gu, X. (2013). A comprehensive review of empirical studies on the washback of language tests over the past two decades. *Foreign Language Testing and Teaching, 2013*(1), 4–17.
Lyu, M. (2015). The exploration and practice of computerized automatic scoring in large-scale English oral test. *China Examinations, 10*, 51–57.
Matoush, M. M., & Fu, D. (2012). Tests of English language as significant thresholds for college-bound Chinese and the washback of test-preparation. *Changing English, 19*(1), 111–121.
Mayring, P. (2004). Qualitative content analysis. In U. Flick, E. von Kardoff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 266–269). Sage.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256.
National Language Commission of the People's Republic of China. (2018). *China's standard of English*. Retrieved from http://www.neea.edu.cn/res/Home/1908/0c960236756 49ac8775ff3422f91a91d.pdf
Prodromou, L. (1995). The backwash effect: From testing to teaching. *ELT Journal, 49*(1), 13–25.
Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317.
Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing, 35*(1), 101–120.
Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1–14). Psychology Press.
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.
Xi, X. (2010). Automated scoring and feedback systems—where are we and where are we heading? *Language Testing, 27*(3), 291–300.
Xi, X. (2012). Validity and the automated scoring of performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 438–451). Routledge.
Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech *using SpeechRater*℠ v1.0. *ETS Research Report Series, 2008*(2), i–102.
Xi, X., Schmidgall, J., & Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English* (pp. 150–175). Palgrave Macmillan.
Yan, K., Hu, G., Wei, S., Dai, L., Li, M., Yang, X., & Feng, G. (2009). Automatic evaluation of English retelling proficiency in large machine-based oral English tests. *Journal of Tsinghua University (sci & Tech), 49*(1), 1356–1362.
Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412.
Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., Zhang, Y., Xu, S., & Fang, L. (2017). Preparing for the speaking tasks of the TOEFL iBT® test: An investigation of the journeys of Chinese test takers. *ETS Research Report Series, 2017*(1), 1–59.
Yu, G., & Jin, Y. (2014). English language assessment in China: Policies, practices and impacts. *Assessment in Education: Principles, Policy & Practice, 21*(3), 245–250.
Zechner, K. (2020). Summary and outlook on automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 192–204). Routledge.
Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S. Y. (2015). Automated scoring of speaking tasks in the test of English-for-teaching (TEFT™). *ETS Research Report Series, 2015*(2), 1–17.
Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*, 883–895.
Zechner, K., & Loukina, A. (2020). Automated scoring of extended spontaneous speech. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice*. CRC Press.

Zhang, Y. (2019). *A study of washback effect of the senior high school entrance oral English testing on junior high school English teaching and learning—taking Zhounan Middle School as an example.* [MA Thesis]. Central China Normal University.

Zhang, M., Bridgeman, B., & Davis, L. (2020). Validity considerations for using automated scoring in speaking assessment. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 21–31). Routledge.

Publisher's Note