Kim, H., & Isaacs, T. (in press). Teachers' voices in the decision to discontinue a public examination reform: Washback effects and implications for utilizing tests as levers for change. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high stakes language testing*. Berlin: Springer.

Hyunjin Kim, Apguejong High School, South Korea.
Email: hk13249@my.bristol.ac.uk

Talia Isaacs, UCL Centre for Applied Linguistics, UCL Institute of Education, University College London.
Email: talia.isaacs@ucl.ac.uk

ABSTRACT

Although a growing awareness of the social nature of assessment has led to an increased interest in washback in language testing, previous research has focused on the effects of existing exams or the introduction of new exams. However, if the introduction or existence of an exam has potential power to produce changes in teaching and learning, withdrawing that exam may also have an impact which deserves our attention. In an attempt to address this gap in washback literature, the present study examined the effects of the decision made by the South Korean Ministry of Education to discontinue the National English Ability Test (NEAT), which had been developed with the intention of promoting curricular change in schools by introducing productive skills assessment in high stakes national testing. This mixed-methods study reports on questionnaire data completed by 72 English teachers from middle schools and high schools in Seoul, Korea and six follow-up interviews to examine how the Ministry's decision affected the instructional practice and perceptions of teachers. Results showed that the abrupt withdrawal of NEAT from being implemented may have had unintended washback effects on the participants' perceptions, if not on their teaching practices. These findings suggest that discontinuing an assessment reform involves more than reverting to the previous state, highlighting the need for greater teacher involvement in the design and implementation of high-stakes assessment policies in order to enhance the potential success of utilizing tests as levers for change.

Teachers' Voices in the Decision to Discontinue a Public Examination Reform: Washback Effects and Implications for Utilizing Tests as Levers for Change

In the field of language testing, increasing recognition of the social nature of assessment has drawn attention to the consequences of testing and the issue of ethics and politics in test use (Alderson & Banjeree, 2001). These considerations now parallel the more traditional concerns about the validation of tests as instruments for measuring language ability. Accordingly, the phenomenon of washback, a widely-used term in applied linguistics denoting the influence of a test on teaching and learning, is now an established area of study in the field. Since the proposal of the list of fifteen 'washback hypotheses' in Alderson and Wall's (1993) seminal article, many empirical studies have been conducted giving evidence to the existence of washback in language testing and shedding light on earlier speculations on the topic (e.g., Wall & Horák, 2011; Xie & Andrews, 2012). However, these accumulating studies have also shown that a wide range of mediating variables, such as the responses of relevant participants, test uses, and educational contexts, are involved in the manifestation of washback (Rea-Dickins & Scott, 2007); thus the broad scope and complex mechanisms of the washback phenomenon have yet to be fully understood.

Despite the complexity and unpredictability involved in washback effects, policy makers have often attempted to draw on the power of tests, both real and perceived, to impose changes in the educational system (Andrews, 2004). One such case in South Korea is the introduction of the National English Ability Test (NEAT), a state-administered English exam that assesses productive skills (speaking and writing) in addition to receptive skills (listening and reading). NEAT for secondary students[1] was developed from 2008 and officially launched in 2012 with the aim of inducing positive washback into English classrooms at middle schools and high schools in Korea. Secondary English education had long been criticized as being restricted by the reading and listening comprehension-based English section of the current college entrance exam (College Scholastic Aptitude Test: CSAT), with NEAT arguably representing an

[1] *NEAT for secondary students* refers to NEAT levels 2 and 3, which were superintended by the Korea Institute of Curriculum and Evaluation (KICE), the organization also in charge of the national college entrance exams. It needs to be distinguished from NEAT level 1 for adults, which is administered by a separate organization, with a different development aim and test use. For the purpose of this chapter, 'NEAT' refers to NEAT levels 2 and 3 for secondary school students and does not encompass NEAT level 1 for adults, which is currently still in implementation.

improvement over the CSAT through its assessment of productive skills (Bachman, 2013; Jin, 2012). Based on the belief that it would be difficult to change school education without adjusting the high-stakes matriculation system, the government planned for NEAT to replace the English section of CSAT completely by 2016. However, such plans became the subject of controversy, as concerns were voiced that the government was pushing ahead with the project too fast, while problems concerning the validity of test items, the objectivity of scoring, and the expertise of raters for productive skills continued to be raised. From a critical applied linguistics perspective, NEAT's reinforcement of native English speaker norms to the exclusion of L2 varieties has been viewed as incongruous with the aim of promoting English language communication in a globalized world, which partially spurred its development (Ahn, 2015). It was also repeatedly pointed out that schools, with large classes made up of students of varying motivation and levels, were not ready to align their education to such a type of assessment and that this would be the source of a surge in private education costs. True to form, private institutes hastened to take advantage of parents' anxiety by using NEAT as a means to promote their programs. Meanwhile, the government delayed making a final decision on adopting NEAT as a substitute for the English section of CSAT, although it had originally planned to make the announcement at the end of 2012 (The National English Ability Test guide, 2011). Finally, following the decision in August 2013 to abandon plans for NEAT to replace the English section of CSAT, the South Korean Ministry of Education (MOE) announced in January 2014 that NEAT would no longer be implemented, indicating a *de facto* termination of the test (Bahk, 2014). Much criticism has been directed towards the government about the billions of South Korean won wasted on this project (Jung & Jung, 2014), but there has been less discussion on the nature of the effects the decision to abolish the test has had on the educational system.

If the introduction of public exams can be used as a force to promote educational reform (Andrews, 2004), it can be postulated that the abolition of such tests may also effect changes, whether positive or negative, intended or unintended. However, it seems that the washback related to such cases of terminating a test has yet to be documented by empirical studies. This research gap has triggered the conceptualization of the current study, along with the first author's professional stake in the changes of high-stakes assessment policy as an in-service teacher in the

research context. This study is broadly grounded in the educational and L2 assessment literature on teachers' mediation of externally-imposed tests with local classroom practice (e.g., Pellegrino, Baxter & Glaser, 1999; Turner, 2012) and aligns with the focus of this book by examining teachers' underrepresented voices in the context of high-stakes testing reform. This literature has highlighted the value of collecting insights from teachers' perceptions of high-stakes assessment practices as key stakeholders in the educational system (East, 2015; Klenowski & Wyatt-Smith, 2012). Winke (2011), for example, argues that teachers' judgments should be taken into account when constructing validity arguments for large-scale tests, suggesting that teachers' perspectives on a test's validity and impact, including its acceptability for making consequential decisions about educational stakeholders, offer 'valuable pieces of information concerning whether tests affect the curriculum as intended' (p. 633). Thus, as part of a larger study on teachers' perspectives on the washback effects and implications of NEAT termination and the future of teaching and assessing productive skills in public English education in South Korea (Kim, 2014), this chapter reports on teachers' perceptions on the washback and implications of NEAT termination to inform ways of mitigating short-lived attempts at implementing assessment reform in the future .

The following research questions were examined:

1. In what ways has the decision not to implement NEAT had an impact on the instructional practice of English teachers at Korean middle schools and high schools?
2. How do the teachers perceive the short-lived NEAT implementation attempt (i.e., backtracking on the proposed change)?

METHOD

Research Design

As the topic of this study is embedded in a complex educational and social context, it seemed that deeper insights could be gained by combining quantitative and qualitative data so that 'words can be used to add meaning to numbers and numbers can be used to add precision to words' (Dörnyei, 2007, p. 45). Therefore, the present study implemented a sequential mixed methods design (Creswell & Plano Clark, 2011), with the first phase (concurrent collection of quantitative and qualitative data) informing the second phase of the study and all evidence being drawn together for

interpretation. Figure 1 shows the nature of the data collected in each phase of the study.

In phase 1, both quantitative and qualitative data were collected by using an online questionnaire, the analysis of which was used to guide and inform the qualitative data collection in Phase 2. In the analysis stage of Phase 1, there was an interplay of quantitative and qualitative data, as the nonnumerical (textual) data from the open-ended questionnaire items were 'quantitized' (i.e., transformed into quantitative data; Tashakkori & Teddlie, 1998) by enumerating the coded categories and converging with other numeric data from the questionnaire to understand trends and develop themes for the follow-up interviews, in addition to being used to validate and clarify the data from the closed-ended questionnaire items. In the final interpretation stage, all data from questionnaires and interviews were integrated with equal weight in the hope that the breadth and depth of each type of data would 'yield results from which one can make better (more accurate) inferences' (Teddlie & Tashakkori, 2009, p. 35).

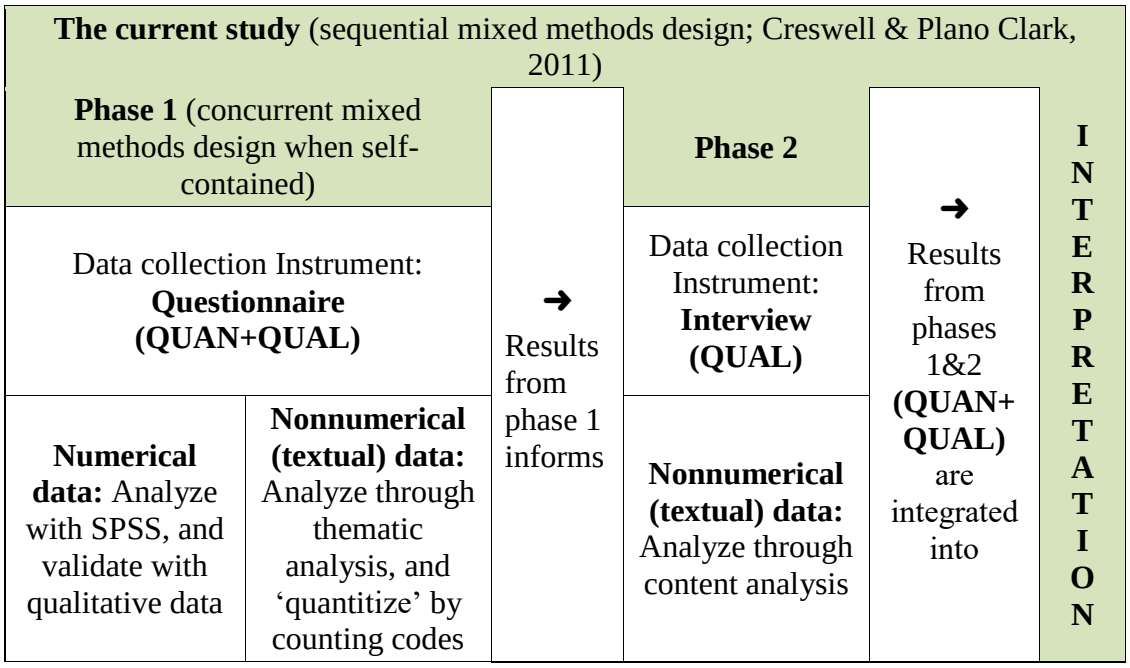

Figure 1. Procedural diagram of the mixed-methods design of the study, linking the nature of the data collected and data analysis techniques to overall data interpretation.

Participants

A combination of convenience and snowball sampling was employed to recruit

participants among English teachers working at middle schools and high schools in South Korea. All were located in Seoul, the South Korean capital and largest city, to minimize the urban-rural and regional differences in the educational context. 21 teachers, who were acquaintances of the first author, were initially contacted as prospective participants. 20 teachers agreed to take part in the research and gave their email addresses. Among these, 14 offered to find other teachers who might be willing to become participants and agreed to disseminate the questionnaire to these additional participants. In the end, completed questionnaires were obtained from 72 participants (61 female, 11 male), which constituted approximately 1.25% of the target population[2], including half from middle schools and half from high school teachers. For the interview, six participants (5 females, 1 male) were selected among the 23 teachers who had volunteered for the follow-up interview in the questionnaire. The choice of interview participants was made based on the questionnaire responses (e.g., one teacher from each cohort who had expressed overall more positive, negative, and neutral responses toward NEAT was selected) to get a balanced perspective on the topic while including an equal ratio of teachers from each school level. Tables 1 and 2 show participant demographics for phases 1 and 2 of the study, respectively, with the data obtained through background questionnaires administered after they had completed the substantive portion of the questionnaire administered in phase 1. In the larger study (Kim, 2014), group differences between middle school and high school teachers in their questionnaire responses examined using a Mann Whitney nonparametric test were nonsignificant ($p > .05$). Therefore, the categories for middle school and high school teachers were collapsed in the questionnaire responses. However, when quoting individual teachers' interview comments verbatim, the following pseudonyms identified whether they taught middle school (M1, M2, M3) or high school (H1, H2, H3).

Table 1. Questionnaire participants' background information (n = 36 middle school and 36 high school teachers**)**

[2] The statistics for the number of English teachers working in secondary schools in Seoul at the time were 5,760 in total with 2,688 middle school teachers and 3,072 high school teachers (Korea Educational Statistics Service, 2014).

Age (years)	≤30	19 (26.4%)
	31–40	31 (43.1%)
	41-50	18 (25.0%)
	≥51	4 (5.6%)
Teaching experience (years)	≤5	17 (23.6%)
	6–10	23 (31.9%)
	11–15	14 (19.4%)
	≥16	18 (25.0%)
Highest level of education	Bachelor's	43 (59.7%)
	Master's	28 (38.9%)
	Doctoral	1 (1.4%)

Table 2. Interview participants' background information (*n* = 3 middle school and 3 high school teachers)

	M1	**M2**	**M3**	**H1**	**H2**	**H3**
Age (years)	31–40	31–40	31–40	≤30	31–40	41–50
Teaching experience (years)	6–10	6–10	11–15	6–10	11–15	6–10
Highest level of education	Bachelor's	Master's	Master's	Bachelor's	Master's	Master's

Data Collection Phase 1 : Questionnaire

A web-based questionnaire was constructed using SurveyMonkey, an online questionnaire builder (http://www.surveymonkey.com). The questionnaire (see Kim, 2014, for the full version of all research instruments) consisted of four sections, each made up of around 6 items. Section A roughly corresponded to RQ1 and sought to examine the impact of NEAT termination on participants' instructional practice, with their instructional practice categorized into the following ten aspects: (1) time allotted to teaching speaking skills, (2) methods for teaching speaking skills, (3) time allotted to teaching writing skills, (4) methods for teaching writing skills, (5) methods for

teaching other skills (e.g., listening, reading, vocabulary, grammar), (6) textbook coverage, (7) supplementary resources,(8) the amount of English used by the teacher, (9) design of performance assessment, (10) test items and format in mid-term and final exams. First, the impact of NEAT implementation on these aspects of instructional practice was examined to serve as a point of comparison. Then, a 4-point Likert scale (not at all influenced, hardly influenced, somewhat influenced, strongly influenced) was used to gauge the effect of NEAT termination on participants' instructional practice in each of the above aspects. Also, two open-ended questions were used to inquire more deeply into the specific nature of the washback (if any) and for aspects where there was no washback, the reason why. Section B corresponded roughly to RQ2 and inquired into teachers' attitudes toward the tested features of NEAT and its implementation through 4-point Likert scales. Teachers were also asked in a dichotomous question whether they supported the MOE's decision to discontinue to NEAT, and to explain their answer in an open-ended question. Finally, their views of the implications of the educational reform attempt were probed in a multiple choice question, with the selection of multiple options made possible. A wide range of both positive and negative options were included based on relevant news articles, blog posts, and suggestions from pilot participants. The order of these options was presented in random using the online tool to control for an order effect. Section C, which probed teachers' opinions on the future direction of teaching and assessing productive skills in public education in South Korea, corresponded to another research question in the larger study (Kim, 2014). Finally, Section D contained questions requesting information about participants' background, reflecting Dörnyei and Taguchi's (2010) suggestion that personal classification questions are best left at the end, so that their sensitive nature would not inhibit the respondents' answers.

The questionnaire was first designed in English and then translated into Korean, the language used in the administration to prevent misunderstandings resulting from using a foreign language. The accuracy of the translation was checked through back-translation by two native speakers of Korean undertaking a Master's program in TESOL/applied linguistics (overall IELTS scores ≥ 6.5) at a research-intensive university in the United Kingdom. The questionnaire was subsequently pilot tested with three English teachers who taught at middle schools and high schools in Gyeongi-do, a neighboring province to Seoul. Revisions were made based on the pilot tests with regard to the order of the questionnaire items, item types, answer options,

and wording. Following piloting, emails with a link to the consent form and final version of the web-based questionnaire were sent out to prospective participants identified through the sampling method outlined in the previous section.

Data Collection Phase 2: Interviews

The second phase of data collection involved the use of semi-structured interviews, in which interview prompts were prepared in advance and administered to each interviewee but the order in which the questions were posed was not pre-determined. This provided the interview with both direction and flexibility. The interview prompts were developed based on the analysis of the questionnaire responses in order to 'illustrate and illuminate questionnaire results' (Dörnyei & Taguchi, 2010, p. 109). Therefore, they included probes about individual responses, especially those that seemed to be open to questions or that digressed from the general trend. For example, according to his/her responses, the interviewees were asked to depict the situation in their classrooms and elaborate how and why NEAT introduction and termination affected various aspects of their teaching practices but not others, or, if there was no influence in all aspects, to describe the extent their teaching practices are influenced by national testing in general. Interviewees were also asked to clarify combinations of answers which could be considered contradictory, such as why they initially supported the introduction of NEAT but later supported its termination, why they rated the direction of changes in NEAT to be desirable but was against NEAT itself, etc. In addition, themes identified in the questionnaires about developing productive skills education in schools, and the interviewee's opinions about the nature of the general trends of answers in the questionnaire were further probed.

Due to practical reasons, the interviews were carried out online. James and Busher (2009) and Salmons (2010) note that there are two forms of communication possible in online interviews: a synchronous mode, which involves interaction in real time and requires both parties to be online at the same time, and an asynchronous mode, which allows for a time lapse between message. Although the flexibility offered by asynchronous communication methods such as email exchanges (Kazmer & Xie, 2008) was not disregarded, it was decided that a synchronous mode of communication would better complement the written questionnaires considering that 'temporal co-presence intensifies online interactions, creating an atmosphere where discussions can flourish'

and that ‘although synchronous communication is written and not spoken, many of its linguistic characteristics mirror the spoken word’ (Stewart & Williams, 2005, p. 405). Among the various types of synchronous communication, two options were considered: video calls through Skype and text-based chats through instant messaging programs. The main difference between the two is the presence of non-verbal cues such as intonation and facial expressions (Opdenakker, 2006). However, it seemed that the nature of the research questions, which involved participants' views about a public matter rather than their personal emotions, did not require a direct observation of paralinguistic cues as information sources. Therefore, it was decided that the use of text-based instant messaging systems as an interview tool should be tested in a set of pilot interviews, one with a pilot participant of the questionnaire and another with an actual questionnaire participant. Both pilot participants agreed that typed exchanges through instant messaging systems offered many advantages to Skype video calls, such as allowing the interviewee to scroll back and re-examine what was said and not having to maintain eye contact while taking time to reflect on his/her answer. On the other hand, it was noted that the absence of non-verbal cues sometimes made it difficult to judge whether the other person had finished talking, and therefore in the actual interviews, a system of typing in “//” at the end of a text chunk was used to signal the end of a message.

The interviews were held over the course of a week, the interview schedule spread out to one each day. Based on each interviewee’s preferred instant messaging program, five interviews were conducted with Kakao Talk PC Version and one with Nateon Messenger, all at a convenient time to avoid distractions. In order to create a comfortable atmosphere in which interviewees could share their thoughts without inhibition, the interviewees were encouraged to follow their usual informal chatting conventions such as the use of emoticons or acronyms. Also, the interviewer was positioned as a fellow teacher listening to a colleague’s views on English education rather than a formal researcher. All interviews were conducted in Korean, as it was the language that both the participants and researcher felt at most ease with.

Data Analysis

The data from the closed-ended questionnaire items were converted into numerical scores using SPSS Version 21 and analyzed using descriptive statistics (means, standard

deviations, frequency counts, distributions). Each response option for a Likert scale was coded as ordinal data ranging from 1 (strongly disagree) to 4 (strongly agree), whereas dichotomous and multiple choice questions were coded as nominal data. For questions that allowed the choice of more than one option, each option was treated as an individual dichotomous question, with the response coded as 1 if chosen by the respondent and 0 if not chosen.

The qualitative data from the open-ended questionnaire items were imported in Microsoft Word and coded by marking segments with representative phrases, which later evolved into categories, using inductive coding roughly adapted from grounded theory procedures (Corbin & Strauss, 2007). The frequencies of the coded categories were counted for each questionnaire item, two of which (25% of the total number of open-ended questions) were checked by a second-coder. Approximately 80% of the coded categories and enumeration were agreed upon; where there was a divergence, consensus was ultimately achieved through discussion. The results of this analysis were combined with the statistically analyzed quantitative data from the closed questions to develop themes to discuss in the follow-up interviews.

While analyzing closed-ended questionnaire data, it was noted that some answers seemed inconsistent. For example, several teachers replied that they had supported the termination of NEAT but were in favor of developing a new test that was similar to NEAT. These cases were identified by examining the combination of answers for the relevant questions, and in case of unexpected combinations, the qualitative data from the mandatory contingent open questions were scrutinized to qualify the result. For the subset of respondents who participated in the follow-up interview, such issues were further probed.

The textual data from the interviews, each stored as separate files by using the ‘save conversation’ function of the instant messaging systems, were organized into a readable format in Microsoft Word for qualitative content analysis (Dörnyei, 2007). The texts were read through several times to gain a general sense of the data. Then, color-coding was used to illuminate informative passages and a preliminary code was assigned to these passages. The initial codes were alliterated and compared to examine whether they could be grouped at a broader level. Throughout this process, reflective memos were used to keep track of emerging groupings of ideas. The resulting themes were linked with the research questions and the analyzed

questionnaire for final interpretation. Interview extracts were then translated into English by the first author for English language research reporting. The final interpretation and translated excerpts included as part of the larger study (Kim, 2014) were shared with the interviewees (i.e., member checking; Seidman, 2006) to confirm the accuracy of the translation and verify that their perspectives had not been misconstrued in interpretation.

RESULTS

Impact of NEAT Termination on Instructional Practice

Figure 2 presents the percentages of responses for the questionnaire item A3, which was designed to investigate teachers' perceptions of the extent of the impact that the decision to terminate NEAT had had on various aspects of their teaching practice. The items are presented in order of the aspects that more teachers marked to be influenced by NEAT termination, computed by adding the percent responses for 'strongly influenced' and 'somewhat influenced'. The top five responses are related to the teaching of productive skills (time allotted in class or teaching methods) and the assessment method frequently used in schools for these skills (performance evaluation).[3]

[3] The final grades that Korean middle school and high school students receive for English are typically based on two types of assessment: *performance evaluation*, which takes the form of direct assessment of tasks in class by the classroom teacher and the paper-based *mid-term and final exams*. Due to the nature of productive skills, they are usually assessed through performance evaluation rather than the mid-term and final exams.

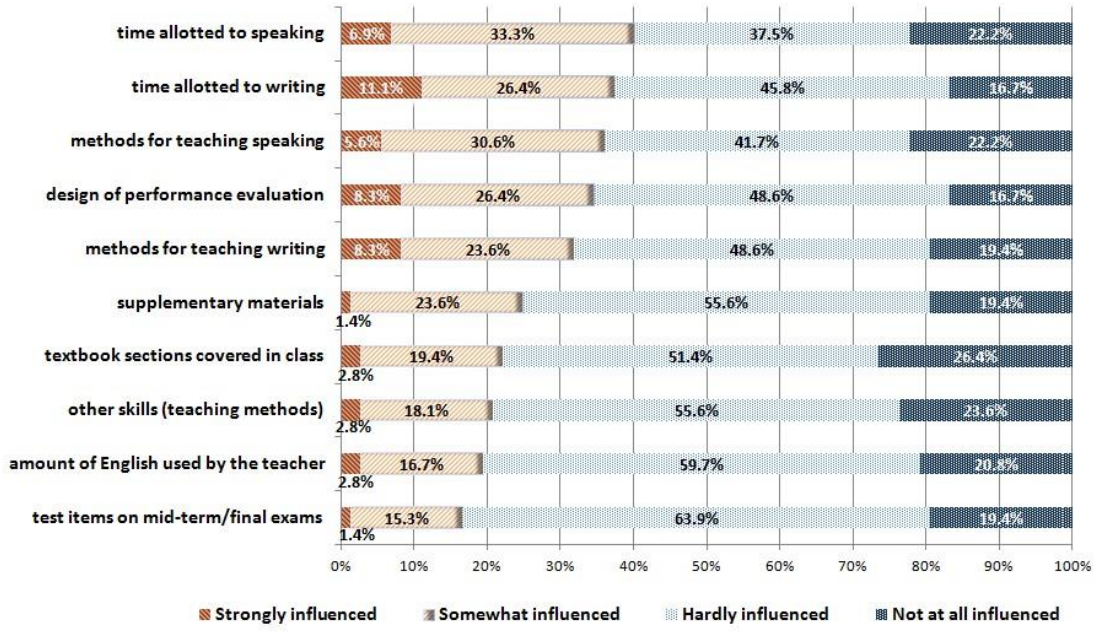

Figure 2. Impact of the termination of NEAT on questionnaire participants' instructional practice

When asked to specify the kind of changes brought about by the decision to discontinue NEAT in an optional open-ended question, 11 (out of the 26 respondents to that item) replied that they would allot less time to writing and speaking, because although important, these skills not immediately assessed in national exams and there is less need for this on the part of the students. Seven respondents replied they perceived change in how they taught; among these, two replied that there would be fewer student-centered activities, two replied that there would be more emphasis on language forms than communication skills, and three replied that they felt less pressure to teach according to NEAT item types. For performance evaluations, seven indicated that the assessment method would become more simplified to more easily scored formats (e.g., closed questions with short answers rather than open-ended essay questions, memorization tasks rather than communicative tasks). Finally, one stated that the general rate of change would slow down, and another intimated that although unsure of any actual change, she felt much less pressure about having to change her traditional teaching methodology.

However, Figure 2 also shows that for all measures obtained, the majority of respondents replied that the decision to discontinue the test would have little or no influence on their teaching practice. When asked to explain why in an optional open-

ended question, 19 (out of the 39 respondents to that item) mentioned that the influence of NEAT's introduction had been minimal, so there was little to be changed by the termination. Among these, nine added that they had doubted the feasibility and full implementation of NEAT from the start. Additionally, 12 responded that they were going to teach in the way they believed was right regardless of the government's testing policy, stating the importance of productive skills. Eight teachers mentioned that they were not influenced by college entrance exams because they worked in middle schools. However, the results for the Mann-Whitney U test to compare answers between middle school and high school teachers for questionnaire item A3 did not show any significant differences for all aspects of teaching practice ($p>.05$), suggesting that this could indicate an attitude towards national testing policies as opposed to an actual difference of impact between school levels within this cohort.

Exploring this further in the interviews, 2 interviewees reported that while the introduction of NEAT had spurred giving more attention to productive skills and attempting innovative teaching methods such as student-centered learning or communicative activities more quickly, the termination of NEAT had no impact on their views about the importance of those skills and, therefore, their teaching practice. Another interviewee expressed that, while the introduction of NEAT in 2009 had initially inspired him to experiment with instructing new things in his classes, the realistic conditions at school had discouraged him from carrying on these efforts long before its termination. He stated:

> **H3**: I tried conducting classes in English and allotting one class per week to English composition. However, I was beset with many barriers, such as the low level of students and lack of time to give guidance to the numerous students. This made me resume my old teaching style, so not much is left for the termination [of NEAT] to make a difference.

While examining the questionnaire results, it was noted that, although the introduction of NEAT had much more influence on the content of teaching in terms of time allotted than on the methods used for teaching productive skills, there was not a large difference on these two aspects in terms of its termination, as is suggested in Figure 3 below:

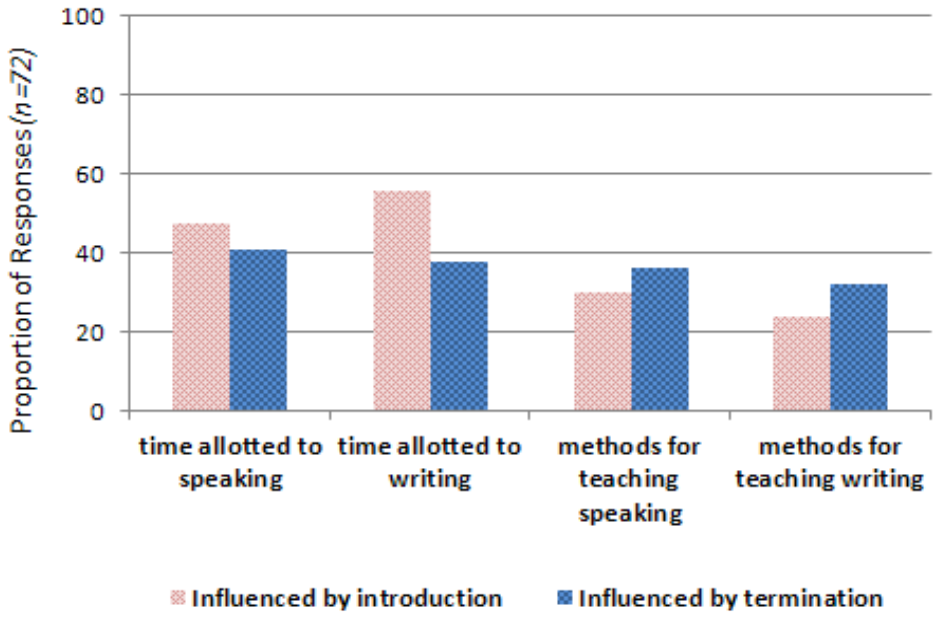

Figure 3. Comparison of the impact of NEAT introduction and termination on questionnaire participants' time allocation and methods for teaching productive skills

This issue was investigated further in the follow-up interviews to confirm whether the responses corresponded to the participants' situations. Answering in the affirmative, the two interviewees who had responded that NEAT termination had impact on the way they taught productive skills but not on the amount of time explained that while the time devoted to each skill remained the same because they recognized the importance of balancing the four skills, they felt less need to tailor their teaching methods to NEAT, which entailed both positive and negative change. One elaborated as follows:

> **M3**: Before, I felt that I should give students practice under NEAT guidelines, whereas now I have more freedom to try other things. Recently, I had students do group work making yoga videos and audio books in English. On the other hand, less pressure also means that I now tend to give low level students only controlled practice or have high level students present their answers without providing them with specific feedback.

Teachers' Perceptions of the NEAT Implementation Attempt

Table 3 summarizes the results of questionnaire item B2, in which respondents were given a table where the differences between CSAT and NEAT were summarized (The National English Ability Test guide, 2011) and asked to rate the appropriateness of each change. Likert scales coded 'very negative (=1)' to 'very positive (=4)' were

analyzed to derive means and standard deviations (SD) for each category. The findings suggest that the majority of teachers had rated the overall direction of change in NEAT favorably, except for the internet-based test delivery and the 4-level classification scoring system. In fact, especially in terms of including productive skills in the assessment, 76.4% of the respondents said that they saw this change as positive.

Table 3. Questionnaire participants' evaluation of the appropriateness of changes to NEAT

CSAT ➜ NEAT levels 2 and 3	**Mean**	**SD**
(1) Skills tested: listening, reading ➜ listening, reading, speaking, writing	2.94	0.837
(2) Question types: multiple choice with five options ➜ listening and reading: multiple choice with four options / speaking and writing: task-based open answers	2.82	0.811
(3) Test delivery: paper-based ➜ internet based	2.36	0.893
(4) Test-taking opportunities: once a year ➜ twice a year	2.81	0.833
(5) Assessment criteria: norm-referenced ➜ criterion-referenced	2.86	0.893
(6) Scoring system: Nine-level classification based on standardized scores ➜ Four-level (A, B, C, F) classification for each skill	2.43	0.853

Note. n = 72, 1=very negative, 2=somewhat negative, 3=somewhat positive, 4=very positive

However, despite positive opinions on the overall direction of change, 75% of the respondents replied that teachers' opinions had not been reflected in the development phase of NEAT (i.e., measures were top down and MOE neglected to take heed of their calls for a more sufficient preparation stage) and almost all respondents (95.8%) replied that teachers had not received enough information about the test. Furthermore, while 47 teachers (65.3%) replied they had been against the introduction of NEAT, 58 (80.6%) replied that they supported the decision to abort it, indicating that 11 teachers had changed their attitudes during the implementation process. Two interviewees fell under this case, and the reasons for their change of attitudes were discussed during the interviews. Teacher H3 explained that the way the exam reform was managed had turned him against it, describing how 'the government seemed to just push ahead with more haste than caution' while he faced many difficulties as a practitioner struggling to incorporate productive skills in the classroom. Teacher H1 replied that the various

problems that remained unresolved during its short implementation, such as controversy over the quality of test items, objectivity of scoring, and expertise of raters, made her think that 'if these problems cannot be solved, it would be better to save time, money, and effort by giving up quickly'. These accounts mirrored the answers to the open-ended questionnaire item B4, in which questionnaire respondents had cited problems to do with scoring (13) [4], the realistic conditions of school education (12), the way of implementation (11), lack of infrastructure (11), validity of test items (8), budget issues (7), the growth of private education costs (6), the time and effort needed on part of relevant stakeholders to prepare for a test based on production, (5), and satisfaction with the current system (5) as reasons for supporting the discontinuation of the test.

In light of the above, perhaps it is not surprising that the most frequently selected options for the implications of the NEAT implementation attempt in questionnaire item B5 (multiple choices possible) were negative ones, as can be seen in Table 4. An analysis of the respondents' answers did reveal that only 14 (19.4%) respondents had chosen only the negative options and that some positive aspects of the attempt were in fact acknowledged as well. However, when asked to rate the NEAT implementation attempt in terms of developing productive skills assessment in public education (questionnaire item B6), 9 respondents (12.5%) gave the rating as very negative, 31 respondents (43%) as negative, 31 respondents (43%), and only 1 respondent (1.3%) as very positive. Therefore, it can be seen that the participants' overall evaluation of the short-lived NEAT implementation attempt, though mixed, leaned toward the negative, even in terms of developing productive skills assessment.

Table 4. Frequencies of questionnaire participants' responses for implications of the NEAT implementation attempt

Rank	Answer Option	Response Count	Response Percent
1	(N) The rash implementation of the test resulted in a waste of national budget and human resources.	53	73.6%

[4] The numbers in brackets refer to the frequency with which the coded category was mentioned in the answers to the open-ended question. 58 participants gave answers about why they supported the termination of the test, and some participants mentioned up to two factors in their responses.

2	(N) The frequent change in national testing policies increased confusion to stakeholders.	49	68.1%
3	(N) The inconsistency of the national testing policy led to mistrust regarding the Ministry of Education's testing policy.	48	66.7%
4	(N) Not being supported by enough research and prior investigation, it led to negative perceptions about national testing policies.	45	62.5%
5	(N) By raising parental anxieties, it resulted in an unnecessary expansion in the market for private education for toddlers, children and elementary school students.	43	59.7%
6	(N) As a top-down innovation that did not take the school context into account, it was a source of demotivation for the teachers.	35	48.6%
7	(P) It raised awareness of the need to develop all four skills in balance.	32	44.4%
8	(P) It helped to identify problems that may arise when assessing productive skills.	31	43.1%
9	(P) It raised awareness of the need to assess productive skills in the public sector.	26	36.1%
10	(N) It resulted in an unnecessary increase in private education for secondary school students.	25	34.7%
11	(N) It increased doubts about the feasibility of assessing productive skills at a national level in the future.	24	33.3%
12	(P) It encouraged teachers to develop expertise in teaching and assessing productive skills.	21	29.2%
13	(P) It helped to make a standard rating scale for speaking and writing assessment.	17	23.6%
14	(P) It contributed to the development of speaking and writing tasks for assessment.	12	16.7%
15	(P) It contributed towards the development of teaching speaking and writing skills such as in the case of the newly revised textbooks.	8	11.1%
16	(P) Through rater training, it was a chance to foster a professional group of raters for speaking and writing assessment tasks that included teachers.	8	11.1%

Note. n = 72, Percentages are rounded to the nearest tenth; (N) stands for negative option and (P) stands for positive option.

DISCUSSION

The focus of this study was the decision not to implement NEAT, a public exam that was developed and introduced in South Korea with the goal of facilitating productive skills teaching in public education, but abandoned before it reached its intended high-

stakes status. In light of the paucity of empirical studies on the impact of abolishing tests, the current study examined the washback effects of the decision to terminate NEAT on the instructional practice of teachers at middle schools and high schools in Seoul, Korea. It also sought to explore these teachers' attitudes toward the attempted educational reform and its implications. In order to meet these aims, both quantitative and qualitative data were collected by means of a questionnaire completed by 72 teachers and follow-up interviews conducted with six teachers. The resulting data were merged to yield findings for this study. Regarding RQ1, it was found that the majority of teachers in this study perceived little or no washback effects in their instructional practice, although a notable proportion did perceive washback effects in terms of teaching and assessing productive skills. For RQ2, teachers' perceptions of the attempted examination reform and its implications were found to be negative, particularly regarding the frequent changes in national testing policy and the rash implementation of the intended reform, which was not supported by sufficient validation research or a thorough investigation into the realistic conditions at schools.

According to the findings, the termination of NEAT seems to have had little or no washback effect on the instructional practice of the majority of teachers participating in this study. The two major reasons the respondents gave for this trend can be related back to the literature on washback. The first was that the respondents had not attached much importance to NEAT during its implementation, so there was a weak initial influence. This is in line with Shohamy et al.'s (1996) observation that the perceived status of the test will influence the degree of washback. In other words, because these teachers had doubted that NEAT would reach its intended high-stakes status due to the controversy surrounding the test, its perceived low status limited the degree of washback they experienced with regard to its implementation and subsequent termination. Another was teachers' belief as a key mediating factor in washback (e.g., Burrows, 2004; Cheng, 2005; Deng & Carless, 2010; Wall, 2005). One-third of the teachers who reported little washback effects of NEAT termination stated their unchanging belief in the importance of productive skills as the reason why the decision to discontinue NEAT had little or no impact on their instructional practice. That is, these teachers felt that the teaching of speaking and writing skills should not be discontinued simply because a high-stakes test focusing on production had been terminated. In other words, their responses were underpinned by the view that

productive skills are important to emphasize in the classroom regardless of the test.

It was also found that the majority of participants held negative views toward the NEAT exam reform attempt, which were based on the arbitrary manner of its implementation rather than on the direction of change itself. Moreover, the closed-ended questionnaire responses that participants most frequently selected as the main implications of the NEAT implementation attempt suggest that its abandonment has led to greater mistrust towards national testing policies. For example, Teacher M2 expressed in the interview that although she could not identify any noticeable washback effects on her teaching practice, her attitude towards national exams had been considerably affected by the short-lived NEAT implementation attempt in that she is now determined not to be swayed by the government's inconsistent assessment policies. Another interviewee (Teacher M3) stated that she supported the development of a new test in the future '*only if* they do not abandon the project again halfway' (emphasis in original). Yet another (Teacher H1) stated that although she had initially welcomed the innovation, its problematic implementation and abrupt withdrawal had now made her skeptical towards another such attempt at exam reform.

Taken together, these results indicate that the precedent of NEAT may have had unintended washback effects on participants' perceptions, if not on their teaching practices. The changes in the participants' attitudes are especially noteworthy in that teachers' perceptions and beliefs play a key role in the realization of change (or lack of change) in instructional practice, both in this study and research to date (e.g., Burrows, 2004; Wall, 2005). Therefore, it is possible that the negative perceptions formed by teachers during the introduction and sudden termination of NEAT may pose an obstacle in similar attempts to use high-stakes state exams as a policy instrument in the future.

To summarize, the results of this study underscore the need for caution when devising policies that employ high-stakes tests as instruments for change and, by reinforcing the findings of washback literature (e.g., Cheng, 2005; Wall, 2005) on the significance of teacher beliefs and attitudes in the manifestation of washback, provides empirical evidence supporting the involvement of teachers in high-stakes testing and policy decision-making. As suggested by Deng and Carless (2010),

changing the direction of high-stakes testing ‘is a helpful first step but cannot be guaranteed to capture the hearts and minds of the teachers’ (p. 300). Moreover, teachers’ reactions to the precedent of NEAT demonstrate how problematic implementation methods of exam reform can further alienate teachers’ attitudes toward national testing policies and possibly reduce teacher engagement in future attempts at using high-stakes language testing as levers for change. If, as described in the previous paragraph, the discontinuation of an exam reform is not simply a matter of turning back the clock and reverting to the previous state, then it should be beneficial to the success of future policies to draw upon the example of NEAT to argue for a greater teacher involvement in the conceptualization and implementation of high-stakes assessment reform.

From the perspective of linking external assessments or assessment reform with classroom practice and curricular goals, including teachers during the development of educational policy and during the process of creating and launching a new test may require more time than top-down measures, but should allow for a more sufficient preparation period that had been lacking in the precedent of NEAT in which to pilot the effects of the intended points of innovation in the actual classroom and possibly improve aspects of test design (Ryan, 2002). Furthermore, the benefits to teachers’ assessment literacy (Gambell, 2004; Taylor, 2009) and the advantages of using insights from teachers in test development and administration would contribute to the overall quality of the test (Winke, 2011; East, 2015), offsetting the added cost of providing professional development training and enlisting teachers’ participation in assessment programs. Teachers’ contribution to assessment practice is also more likely produce a stronger sense of ownership in exam reform, which may lead to a heighted trust in its feasibility and an elevation of the exam’s perceived status on part of teachers. The case of NEAT attests to Klenowski and Wyatt-Smith’s (2012) argument that teachers, rather than the test itself, is the key agent of change in educational policies. Therefore, it appears that increasing the role of teachers at all stages of exam reform would strengthen the potential of future attempts to effect change in the classroom through high-stakes testing.

CONCLUSION

This mixed-methods study documents teachers’ voices on the termination of a public

exam that was part of a top-down approach to bring about changes in English language education through changes in large-scale national testing. It must be acknowledged that as a small-scale study, the results can only represent the situation and views of the 72 teachers who participated in this study. All participants worked at schools in Seoul, which has distinct regional characteristics as a metropolitan city. Therefore, the findings of this study may not be generalized to all middle school and high school teachers in Korea. Furthermore, as it is difficult to measure the accuracy of the participants' self-report data, the results can only reflect the *perceived* impact of and implications of the termination of NEAT. Finally, the fact that the test in question had been aborted in its initiation stage made it difficult to arrive at definite conclusions about the washback effects of terminating a test.

Notwithstanding these limitations, this study is unique in its focus on the washback effects of discontinuing an exam. To date, washback has mainly been examined in relation to existing exams or introduction of exams, although test termination can also be seen as a change in language testing that could function as a *de facto* policy and generate washback effects unintended and unpredicted by those who devised it. Therefore, expanding the boundaries of washback to include cases of test termination in diverse national contexts can contribute to our understanding of the phenomenon and better inform future assessment policies.

The results of this study confirm previous findings from the washback literature which identify teachers' beliefs and attitudes as a pivotal factor in the realization of educational change and suggest that arbitrary implementation, regardless of the direction of change, can negatively influence their perceptions, which in turn may become an impediment in similar attempts at reform in the future. These findings give strong support for the benefit of involving teachers in test development and implementation, not least taking their voices into account when generating new educational policy even at conceptualization phase. Although the present study is grounded in the specific context of the abrupt withdrawal of a state-administered national exam in South Korea, its implications of the need for greater teacher involvement in language testing policies is applicable to all contexts where the exclusion of teachers from top-down educational reform poses a threat to the realization of the intended change in the classroom. Indeed, as key educational stakeholders who are often also the major implementers of new teaching, learning and

assessment endeavours and who are also likely to liaise with other major stakeholders (e.g., parents, principals, and learners themselves), giving sufficient attention to teachers' voices and eliciting their participation in further undertakings of educational reform (i.e., through facilitating dialogue at all stages of the process, ensuring adequate time for professional development training and to adapt classroom conditions to the new system, etc.) would appear to be crucial for successful utilization of high-stakes testing as lever for change.

Acknowledgments

This study was conducted in partial fulfillment of the first author's Master's degree in TESOL/Applied Linguistics at the University of Bristol in the United Kingdom.

REFERENCES

Ahn, H. (2015). Assessing proficiency in the National English Ability Test (NEAT) in South Korea. *English Today, 31,* 34-42

Alderson, J. C. & Banerjee, J. (2001). Language testing and assessment (part one). *Language Teaching, 34,* 213-236

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

Andrews, S. (2004). Washback and curriculum innovation. (In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-50). Mahwah, NJ: Lawrence Erlbaum)

Bachman, L. F. (2013, April 10). NEAT to have positive impact on English learning. *Korean Herald*. Retrieved June 2, 2014, from http://www.koreatimes.co.kr/www/news/nation/2013/10/181_133721.html

Bahk, E. (2014, January 15). Homegrown English test to be phased out. *The Korea*

Times. Retrieved February 2, 2014, from http://koreatimes.co.kr/www/news/nation/2014/01/113_149859.html

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. (In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing: Research contexts and methods* (pp. 113-128). Mahwah, NJ: Lawrence Erlbaum)

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

Corbin, J. M. & Strauss, A. (2007). *Basics of qualitative research techniques and procedures for developing grounded theory* (3rd ed.). (Thousand Oaks, CA: SAGE)

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). (Thousand Oaks, CA: SAGE)

Deng, C. & Carless, D. R. (2010). Examination preparation or effective teaching: Conflicting priorities in the implementation of a pedagogic innovation. *Language Assessment Quarterly, 7*, 285-302

Dörnyei, Z. (2007). *Research methods in applied linguistics.* (Oxford: Oxford University Press)

Dörnyei, Z. & Taguchi, T. (2010). *: Construction, administration, and processing* (2nd ed.). (London: Routledge)

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing, 32*, 101-120

Gambell, T. (2004). Teachers working around large-scale assessment: Reconstructing professionalism and professional development. *English Teaching: Practice and Critique*, 3(2), 48-73

James, N. & Busher, H. (2009). *Online interviewing.* (Thousand Oaks, CA: SAGE)

Jin, K. (2012, September 25). Will new English test facilitate communication in the classroom? *Korean Herald*. Retrieved June 2, 2014, from http://www.koreaherald.com/view.php?ud=20120925000720

Jung, M. & Jung, S. (2014, May 21). Questions remain over billion blown on NEAT. *The Korea Times*. Retrieved May 29, 2014, from http://www.koreatimes.co.kr/www/news/nation/2014/08/181_157589.html

Kazmer, M. & Xie, B. (2008). Qualitative interviewing in internet studies: Playing with the media, playing with the method. *Information, Community and Society, 11*, 257-278

Kim, H. (2014). Teachers' voices in the decision to discontinue a public examination reform: Washback effects and implications for teaching and assessing productive skills of English in Korean secondary schools. Unpublished Master's Dissertation, University of Bristol, UK

Klenowski, V., & Wyatt-Smith, C. (2012). The impact of high-stakes testing: The Australian story. *Assessment in Education: Principles, Policy & Practice, 19*(1), 65-79

Korea Educational Statistics Service (2014).*Yoochojoongdeung kyoyuk tongye* [*Educational statistics for kindergartens, elementary schools, and secondary schools*]. Retrieved May 21, 2014, from Korea Educational Statistics Service Web site http://kess.kedi.re.kr/stats/intro?menuCd=0101&survSeq=2014&itemCode=01

Opdenakker, R. (2006). Advantages and disadvantages of four interview techniques in qualitative research. *Forum: Qualitative Social Research*, *7*. Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/175/392.do

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, *24*, 307-353.

Rea-Dickins, P. & Scott, C. (2007). Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Assessment in Education: Principles, Policy and Practice, 14*, 1-7.

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21*(1), 7-15.

Salmons, J. (2010). *Online interviews in real time*. (Thousand Oaks, CA: SAGE)

Seidman, I. (2006). *Interviewing as qualitative research.* (3rd ed.). (New York: Teachers College Press)

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

Stewart, K. & Williams, M. (2005). Researching online populations: The use of online focus groups for social research. *Qualitative Research, 5*, 395-416

Tashakkori, A. & Teddlie, C. B. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. (Thousand Oaks, CA: SAGE)

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29(1), 21–36.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research*. (Thousand Oaks, CA: SAGE)

Turner, C. E. (2012). Classroom assessment. (In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 65-78). New York: Taylor & Francis)

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. (Cambridge: Cambridge University Press)

Wall, D. & Horák, T. (2011). *The impact of changes in the TOEFL examination on teaching in a sample of countries in Europe: Phase 3, the role of the coursebook and phase 4, describing change*. TOEFL iBT Research Report TOEFL iBT-17. (Princeton, NJ: ETS)

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly, 45*(4), 628-660

Xie, Q. & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*, 49-70