This is an Author Accepted Manuscript of an article published in *Language Assessment Quarterly*, 8 March 2021, © Taylor and Francis.

Rossi, O., & Brunfaut, T. (2021). Text authenticity in listening assessment: Can item writers be trained to produce authentic-sounding texts? *Language Assessment Quarterly*.

Available online: https://doi.org/10.1080/15434303.2021.1895162

Text authenticity in listening assessment: Can item writers be trained to produce authentic-sounding texts?

Olena Rossi and Tineke Brunfaut

Lancaster University, UK

Abstract

A long-standing debate in the testing of listening concerns the authenticity of the listening input. On the one hand, listening texts produced by item-writers often lack spoken language characteristics. On the other hand, real-life recordings are often too context-specific to stand alone, or not suitable for item generation. In this study, we explored the effectiveness of an existing item-writing training course to produce authentic-sounding listening texts within the constraints of test specifications. Twenty-five trainees took an online item-writing course including training on creating authentic-sounding listening texts. Prior to and after the course, they developed a listening task. The resulting listening texts were judged on authenticity by three professional item reviewers and analysed linguistically by the researchers. Additionally, we interviewed the trainees following each item-writing event and analysed their online discussions from during the course. Statistical comparison of the pre-and post-course authenticity scores revealed a positive effect of the training on item-writers' ability to produce authentic-sounding listening texts, while the linguistic analysis demonstrated that the texts produced after the training contained more instances of spoken language. The interviews and discussions revealed that item-writers' awareness of spoken language features and their text production techniques influenced their ability to develop authentic-sounding texts.

Introduction

The concept of authenticity was originally put forward by language teaching researchers as a reaction to early English language teaching textbooks which were characterised by stilted, artificially-sounding texts. Authentic language was thereby understood as "a stretch of real language, produced by a real speaker or writer for a real audience" (Morrow, 1977, p.13). This view was enthusiastically adopted by the field of language assessment, with authentic tests aiming to use and elicit "the language used by real people in real life" (Shohamy & Reves, 1985, p.57). However, proponents soon realized the challenges of putting it into practice (Spolsky, 1985; Shohamy and Reves, 1985). Consequently, Bachman (1990) proposed an alternative notion of test authenticity based on work by Widdowson (1979) who differentiated between *authenticity* and *genuineness*. Language produced for real-life purposes is seen as *genuine*, however the use of genuine texts in teaching or testing does not automatically imply authenticity if other components of authentic use are missing, such as a shared context and "knowledge of conventions" (Widdowson, 1979, p.166). Recent conceptualisations of language authenticity see it as a continuum – with the test designer deciding on how to position a test on the continuum depending on multiple considerations – rather than as a dichotomy of authentic/inauthentic, since that inadvertently leads to value judgements (see e.g., Pinner, 2014).

In language testing, debates around authenticity are particularly prominent in relation to listening assessment, with varied views on whether the use of genuine texts is desirable, necessary, or indeed possible. While there is a solid body of research on the qualities of genuine spoken texts, in particular involving corpus methods (e.g. Biber et al., 2004; Carter & McCarthy, 1997), limited empirical research has specifically looked into issues regarding listening text authenticity and item writing. Therefore, in this study we explore whether item writers can be trained to create authentic-sounding listening texts (not items/tasks), regardless of their origin.

Literature review

Text authenticity in listening assessment

In the context of testing listening, criticism often focuses on the lack of use of genuine listening texts: "the testing of listening… still involves candidates being played an audio tape specially constructed for the test and recorded by actors whose disembodied voices boom out in the examination hall" (White, 2018, p.1). Instead, the use of genuine listening input is argued to be an ideal testing choice, with such texts "provid[ing] better linguistic models" (Buendgens-Kosten, 2014, p.458) because they possess features of spoken language; are not graded, so can better predict what

learners will be able to cope with in real-life situations (Field, 2019a); and their use in tests will lead to positive washback in the classroom (Wagner & Toth, 2014). Within such views, it is argued that the use of genuine listening texts in testing automatically triggers listening processes characteristic of the target language use (TLU) domain (Field, 2013) and "ensure[s] that the communicative competence of test takers is actually being assessed" (Wagner, 2016, p.121), thus leading to valid assessment. Also, stakeholders often regard the use of genuine listening input as a highly credible approach to testing listening, resulting in high face validity.

However, following Widdowson's (1979) genuine-authentic distinction, a text is not inherently authentic just because it is a sample of 'real language'. Used in a test, a genuine listening text might lose its authenticity, firstly, because the relevant context and intended audience are missing and it no longer meets its rhetoric and interactional purposes. MacDonald, Badger and White (2000), for example, found that the use of genuine university lectures, instead of constructed ones, did not bring a positive effect on EAP material authenticity because genuine lecture extracts "only partially replicate the features of discourse and language found in the target situation" (p.264). Similarly, experienced item writers interviewed by Salisbury (2005) reported that "using genuine oral texts would be unworkable for listening tests because … they depend so firmly on shared contextual understanding which cannot be provided in test conditions" (p.165).

Secondly, a genuine listening text used in a test might become inauthenticated because in real-life situations there is often no task following a listening event (Buck, 2001). Genuine listening texts may not always lend themselves to being accompanied by tasks and items. If the concern is not just face validity, but also construct validity, the cognitive processes that underlie listening processing in the TLU domain should be defined, and listening tests should aim to activate those processes through carefully constructed tasks based on suitable texts – which does not necessarily mean genuine texts. Therefore, several scholars (e.g., Buck, 2018; Lynch, 2009) argued that it is acceptable to modify a genuine text if this is needed to create better items. Similarly, it is considered acceptable to use 'realistic texts', i.e. texts constructed for testing purposes which feature spoken language characteristics.

Genuine texts are in fact hardly used in language testing (Wagner, 2016), due to reasons such as the difficulty of obtaining copyright permissions, the unprofessional sound quality of genuine recordings, or the fact that detailed and elaborate item specifications make it impossible to find genuine texts that fit the specs (Gilmore, 2007; Wagner, 2014; Field, 2019b). The latter points towards the text-item tension at the heart of the listening authenticity debate. Richards (2007) argued that the lack of clarity and specificity in genuine conversations makes them virtually unusable

for testing purposes. Although an advocate of genuine texts, Field (2019a) also acknowledged that "much authentic interactional speech may be uninteresting or very loosely structured" (p.56), which essentially leaves us with speeches and scripted genuine texts, such as TV and radio programmes, to select from. However, even those often have to be "doctored" (Widdowson, 2003, p.105) to suit testing purposes. Field (2013) provided an account of Item-Writing Guidelines for Cambridge ESOL tests (now Cambridge Assessment English): although the use of genuine texts is encouraged, they have to be modified to "ensure that items are spaced evenly throughout text", "ensure that the piece has a clear introduction", and "add distraction to a text" (p.111). When multiple parallel test versions are required, the use of genuine texts becomes even more problematic because of "uncontrolled variations" (Green, 2014, p. 12) in genuine text characteristics. These and many other considerations (e.g., low information density in many genuine texts, which hinders the creation of enough items) often make text modification inevitable. Of course, the above-mentioned challenges regarding the use of genuine listening texts do not mean they are not useful in listening test design. Rather, it is important to be aware that listening text genuineness does *not automatically* mean such texts are always the best testing choice.

Alternatives to using genuine texts in testing listening

If genuine texts are often unsuitable for testing, what other options exist? Alderson et al. (2006) described two alternative types of input texts: adapted/simplified, and pedagogic (i.e. specifically created for testing purposes). Field (2019a) expressed a preference for text *adaptation* whereby item writers modify "transcripts of well-sourced material" (p.57) to fit item specifications. This approach, however, might lead to the loss of original text features as item writers' attention shifts from text authenticity to conforming to task specifications. Field warned that care should be taken to preserve spoken features of the original texts. However, as acknowledged by Green (2014), "adapting texts in ways that are sympathetic both to the original … and the test developer's purposes is a very demanding task, even for the most experienced item writers" (p.12).

'Pedagogic' input text creation can be approached in different ways. Item writers might be commissioned to produce written scripts which are then studio-recorded by voice actors – a technique we define here as *scripting*. Scripting is the most frequently used technique of listening text development; however, scripted texts are often criticized for lacking spoken language characteristics: they typically have written-like grammar; they usually lack oral discursive features such as hesitations, back-channelling, false starts, pauses, and repetitions; their vocabulary is not colloquial enough; and they are recorded by trained actors who enunciate clearly and thus they lack connected speech (Wagner & Ockey, 2018). Field (2019a) argued that the orality of scripted texts is

"heavily dependent upon how sensitive the item writer's ear is to the content and cadences of natural speech" (p.56). This sensitivity, however, might potentially be developed in item writers, who could then be trained to produce oral texts which more fully bear the characteristics of spoken language and at the same time avoid the above-mentioned problems of genuine texts (Buck, 2001; Lynch, 2009; Richards, 2007). In fact, Field (2019b) formulated recommendations for item writers on how to increase the authenticity of scripted texts. Wagner (2018), as an improvement on the *scripting* technique, experimented with purposefully adding spoken language features to scripted listening texts lacking oral features. He called this technique *text authentication*. He found that it was possible to incorporate several spoken language features, such as false starts, reformulations, and back-channelling. However, discoursal and organizational patterns were more problematic to replicate.

Another approach to listening text development involves "decid[ing] the content in advance, but only the ideas, not the words. Speakers then speak freely expressing these ideas in whatever way comes naturally" (Buck, 2001, p.163). Buck called such text development from content points *semi-scripting*. Clark (2014), for example, who aimed to produce listening texts containing features of academic lectures but "shorter in length and more appropriate for a lay audience" (p.8), studio-recorded subject specialists speaking from prepared notes. It was noted, however, that using just any native/highly-proficient speaker for creating semi-scripted texts might be problematic as they are not trained in conforming to item specifications. Therefore, Green (2017) suggested that item writers should act as speakers during text development: they record themselves speaking and then transcribe the recorded speech capturing spoken language characteristics. The transcribed text is later studio-recorded by voice actors who should be provided with detailed guidance on how to voice the text, for example by using spoken language transcription conventions similar to those in discourse analysis studies (Field, 2019b).

One more technique for listening text production is *improvising* whereby speakers are "asked to follow certain role-play guidelines" (Field, 2019a, p.55) but otherwise can talk freely. Wagner and Toth (2014) used this technique to record conversations by native speaker volunteers who were given instructions on the topic and situation only. One problem with improvising might be in employing speakers uninitiated to item-writing and thus not trained in conforming to item/task specifications, especially as improvisation is less structured then semi-scripting because speakers are not guided by pre-defined content points to be tested in items.[1]

[1] Ambiguity exists in the literature regarding listening text development terminology. What Alderson et al. (2006) call 'adapted/simplified' texts, Field (2019a) terms 'semi-scripted' – a name that Buck (2001) uses for

In summary, in testing listening, text authenticity can be argued to lie in texts that are perceived or judged to reflect 'real-life' spoken text purposes and characteristics, regardless of their origin (since genuine texts can also become inauthenticated, as explained above). We call such texts 'authentic-sounding' (see Thorn (2018) for similar understanding of the term). In practice, in addition to genuine texts, there are several options for listening input text development: genuine text adaptation, improvising, semi-scripting, and scripting. Each approach has weaknesses and it is important to consider the test purpose and associated with it the criticality and feasibility of authenticity (e.g. compare proficiency with diagnostic purposes), in determining the approach taken.

Irrespective of the approach, ultimately the success of an approach will be mediated by the skills of the item writers involved. To our knowledge, empirical investigations of the possibility and effectiveness of training item writers to produce authentic-sounding texts (whether they are adapted from genuine, improvised, semi-scripted, or scripted) are lacking. To begin to address this gap, this study aimed to investigate the following research questions:

1. To what extent do item-writer trainees produce authentic-sounding listening texts, as evaluated by professional item reviewers, prior to vs. after taking an item-writing course?
2. To what extent do the listening texts produced by item-writer trainees pre-training vs. post-training demonstrate spoken language characteristics?
3. How do item-writer trainees' insights into and approaches to the production of authentic-sounding listening texts evolve from before, during, to upon completion of an item-writing course? What aspects of the training do the trainees perceive as beneficial?

Methodology

Item-writing training course

This study explored the effect of an item writer training course conducted for employees of a large-scale language testing organisation. The training aimed to increase participants' theoretical knowledge and practical skills in writing a range of items for all language skills. The course was delivered asynchronously online and included theoretical input, group discussion activities, and item-writing practice. It consisted of six modules, each focusing on a particular language skill and

another text development technique. What Field (2019a) calls 'improvised' texts, Wagner and Toth (2014) call 'unscripted'. Throughout this article, the terms 'adapted', 'improvised', 'semi-scripted', 'scripted', and 'authenticated' are used as defined in this section.

running for two weeks each time. For each module, participants were divided into small groups to discuss module tasks and give feedback on each other's items. The course was taught by two tutors experienced in item writing, reviewing, and online training: the main tutor, who had also developed all course materials, and an assistant tutor who, together with the main tutor, monitored group discussions and provided feedback on items written by participants.

Module 6 of the training was dedicated to item-writing for testing listening. The theme 'authenticity of listening input' was given special consideration. In week 1, participants studied a lecture recorded by the course tutor, covering the construct of listening and higher- and lower-level listening processes (Field, 2013), what makes listening difficult (Green, 2017; also see Bloomfield et al., 2010), and differences between spoken and written language including phonological, lexical, grammatical, and discursive characteristics of spoken texts (Carter & McCarthy, 2007; Wagner, 2016). In another recorded lecture, participants were introduced to practical techniques for developing listening input texts. These included the 'textmapping' technique of exploiting genuine sound files (Green, 2017), and techniques for achieving authenticity in item writer created texts: semi-scripting (Buck, 2001) and introducing spoken language characteristics into scripted texts (cf. Wagner, 2018). Following the lectures, participants completed two practical tasks: 1) 'textmapping' a genuine sound file and, through a group discussion activity, arriving at a consensus about the file's salient content points to be targeted in items; 2) reflecting on the authenticity of a listening text each participant had developed as part of a pre-course assignment, and revising the texts to make them more authentic-sounding. Participants posted their improved texts in their discussion groups for peer-feedback. During week 2, participants were introduced to principles and techniques of producing listening test items in another recorded lecture. Each participant then developed three listening tasks, including texts and items, according to a set of specifications provided by the tutors. Participants discussed their tasks in groups with an opportunity to revise them before submission to the course tutors, who provided feedback, including on authenticity.

Item-writing trainees

Twenty-five participants took part in the course, six female and 19 male. Their ages ranged between 29-60 years old (M=40.4). Except for one Polish-L1 and one Dutch-L1 speaker, all were native speakers of English (British, American, Australian and New Zealand). All participants were educated to a minimum of BA-level and were qualified teachers of English as a Second/Foreign Language; they all held a CELTA, seven also a DELTA, and two a Postgraduate Certificate in Education. They had 3-17 years' ESL/EFL teaching experience (M=8.5). All were professional examiners. None had previous

experience writing items for a professional exam board, although they had some experience writing tests for classroom use or local university entry. Below, pseudonyms are used for individual trainees.

Data collection

To be able to investigate the effect of the item-writing training, participants completed an assignment prior to the start of the course. This assignment consisted of four item development tasks, one of which concerned writing a CEFR-B1 general English listening task which we focus on here (the others were grammar and writing tasks). For the listening task, the participants were required to develop a written-out listening input text and six gap-fill items to go with it. To this end, the trainees were provided with a set of specifications (see Appendix 1), with the specifications' design based on the socio-cognitive framework for test development and validation (Weir, 2005). The required listening text, as stated in the specifications, was a monologue of up to 300 words in one of the following genres: a lecture/presentation, a radio programme, or a short talk. The specs stipulated that the text should be authentic-sounding for the chosen genre but did not impose restrictions on how or where the listening input could be sourced. After a trainee submitted their pre-course assignment, a semi-structured interview was conducted with 17 willing participants via voice call facility. In this interview, the trainee was invited to reflect on his/her approach to producing the four test tasks, including the listening task. More specifically, the trainee was prompted to elaborate on how they had gone about creating the tasks, the resources they had used to do so, and any difficulties they had encountered; they were not specifically prompted to comment on listening text authenticity.

Following the course, participants completed a post-course item-writing assignment identical to the pre-course one described above, using the same specs, but producing new listening tasks rather than simply revising their pre-course ones. They were also interviewed again after submitting their post-course assignment. The same interview schedule as pre-course was used, but with the addition that trainees were encouraged to comment on the helpfulness of the training (or lack thereof) for writing the tasks. Nineteen trainees made themselves available for the post-course interviews (16 of which had also participated in the pre-course interviews).

Data analysis

To evaluate the quality of the items developed by the trainees prior to vs. after the course, first an evaluation scale was developed. The scale consisted of a set of criteria based on the item requirements stipulated in the specifications. Three professional item-reviewers were then recruited to rate the quality of the pre- and post-course item-writing tasks using this evaluation scale. The

reviewers had 5-11 years' experience (M= 8 years) in reviewing items for a range of international language test organisations and a range of standardised English proficiency tests. Two were male, one was female. Each reviewer individually judged the quality of the items. It should hereby be noted that, prior to item quality evaluation, trainees' pre- and post-course tasks were anonymized and randomised so that the item reviewers would not know who had developed the listening task and whether it was from before or after the course.

For the listening item-writing task under focus here, we followed an item evaluation approach which is also sometimes used for item development in operational language testing contexts. The scale developed to evaluate participants' listening tasks comprised 21 criteria, in line with the different requirements stated in the listening specs. This included one criterion on listening text authenticity, which the item reviewers were asked to score on a three-band scale: '2' – the text sounds fully authentic according to the genre; '1' – the text sounds mostly authentic according to the genre, while some minor parts do not; '0' – the text sounds inauthentic according to the genre. This 3-band scale reflects the decisions made during item review in operational contexts: accept the item, return it for further revision, or reject it. Agreement between the three reviewers was established using the intraclass correlation coefficient (ICC) two-way mixed-effect model based on absolute agreement. For the listening text authenticity scores, this was .58 (fair agreement – Cichetti, 1994). For comparison, in a similar study using three judges by O'Neill et al. (2019) the agreement was .47. It should also be noted that low agreement is typical for studies where the expert judgement method is used: it was found that from 20 (Bejar, 1983) to 31 judges (O'Neill et al., 2019) have to be employed to have a robust exact agreement, which is rarely feasible. However, expert judgement (item review) is used as a standard method of assessing item quality in operational testing. For this study, median scores were used to establish so-called 'final item evaluation' scores that would best reflect the opinions of all three reviewers together.

Then, to explore to what extent item writer trainees were able to produce authentic-sounding listening texts prior to vs. after the training as evaluated by professional item reviewers (RQ1), the evaluation scores on the authenticity criterion for trainees' pre-course listening text were compared with the scores for their post-course listening text. This was done by means of Wilcoxon signed-rank tests.

To establish to what extent the listening texts produced by course participants pre-training vs. post-training demonstrated spoken language characteristics (RQ2), linguistic analyses were run separately on the set of pre-course vs. the set of post-course listening texts produced by the trainees. The texts were analysed in terms of their lexical, grammatical and discursive characteristics.

The specific characteristics focused on (see Table 3) were selected on the basis of prior research into linguistic differences between spoken and written language (Carter & McCarthy, 2007) and studies that examined the issue of authenticity in listening texts (Gilmore, 2004; Wagner, 2018), thereby taking into account the text genres targeted in the present study. Depending on the linguistic feature, the texts were coded manually (e.g. to establish the number of subordinate clauses) or using the LancsBox corpus toolbox (e.g. to identify spoken discourse markers) (Brezina et al., 2018). These analyses were conducted by one of the researchers, with 20% of the texts double-coded by another coder. The average coder agreement was 89%, with the disagreements occurring almost exclusively on two codes only. The two coders discussed the cases of disagreement until they reached a consensus. All texts were then recoded on these two codes by the first coder.

Finally, to gain insights into how item-writer trainees' perceptions of and approaches to the production of authentic-sounding listening texts evolved from before, during, to upon completion of the item-writing course, as well as the aspects of the training that the trainees perceived as beneficial (RQ3), two types of qualitative data were analysed: the pre- and post-course interviews with the trainees, and the group discussions they had as part of the course's module on 'Testing Listening'. First, the interview recordings were transcribed, and the group discussions (which had been conducted online in writing during the course) were downloaded. Second, we identified those parts where the participants talked about listening text authenticity. Then, through a process of multiple readings, we arrived at 16 thematic codes and coded the datasets accordingly in ATLAS.ti. Finally, we identified overarching themes across the codes: for the pre-course interviews they were 'attitude to authenticity' and 'inauthentic text features'; for the post-course interviews, they were 'approach to text development', 'spoken language features', 'difficulties in producing authentic-sounding texts', and 'role of training'; for the group discussions, they were 'spoken language features', 'authenticity and text genre', and 'peer feedback on text authenticity'.

Results

Item reviewers' judgements

Wilcoxon signed-rank tests indicated that the post-course ranks for the text authenticity criterion were statistically significantly higher than the pre-course ranks, with medium to large effect sizes (see Table 1). These results hold for each individual reviewer's evaluation, as well as the combined median scores. This suggests that the training had a positive effect on trainees' ability to develop authentic listening texts and that item writers can be trained to produce authentic-sounding texts.

Table 1: Wilcoxon signed-rank test results

	Reviewer 1	Reviewer 2	Reviewer 3	Reviewers' median
Z-score	-2.841	-1.999	-3.532	-3.071
Asymp.Sig. (2-tailed)	0.005	0.046	0.000	0.002
Effect size	0.40	0.28	0.50	0.43

A closer look at individual participant scores, however, revealed that the training effect was not uniform (see Table 2). Three trainee profiles were identified:

a) eight trainees whose pre-training listening texts gained low scores on the authenticity criterion (pre-course score '0' or '1'), but who fully developed their ability to produce authentic-sounding texts following the course (post-course score '2');

b) fifteen trainees whose ability to produce authentic-sounding texts still had scope for further improvement following the course (pre-course score '0' or '1'; post-course score still '1');

c) two trainees whose scores were already maximally high pre-course so there was no room for post-course improvement (pre- and post-course score '2').

Table 2: Text authenticity criterion: comparison of pre- to post-course scores

Trainee profile	Number of trainees	Trainees	Pre-course score	Post-course score
a	3	Chloe, Henry, Ted	0	2
	5	Adam, Emily, James, Lucas, Mathew	1	2
b	3	Arthur, Joe, Rose	0	1
	12	Alex, Austin, Daniel, Jake, Josh, Liz, Lucy, Luke, Mason, Nathan, Olivia, Stanley	1	1
c	2	Logan, Ryan	2	2

Text analysis

Next, the spoken language characteristics of the listening texts produced pre- and post-course were analysed. Because the cumulative length of the pre- and post-course texts was unequal, the number of instances per variable is presented as a ratio to allow for valid comparisons (*Table 3*). Such presentation, however, might prove misleading. First, it should be kept in mind that low percentages are expected for many variables, e.g. a spoken text will never consist exclusively (100%) or primarily of interjections, ellipsis, repetitions, etc. Second, low percentages for some variables conceal large differences in the actual number of instances when comparing before with after the training; that is why the size of the change is indicated in the last column of the table.

Table 3: Listening text characteristics

Variable	**Pre-course**	**Post-course**	**Percentage change**
Text length (mean)	226 words	280 words	+24%
Subordinate clauses (per clause)	27.3%	23.5%	-14%
Embedded clauses (per clause)	1.9%	0.3%	-84%
Simple linking words (per word)	2.1%	2.7%	+28%
Exclamations (per clause)	0.5%	5.8%	+1,060%
Questions (per clause)	1.2%	2.3%	+92%
1st person utterances	41%	48%	+17%
Passive verb forms (per word)	0.6%	0.3%	-50%
Spoken discourse markers (per word)	0.52%	1.49%	+186%
Interjections (per word)	0.05%	0.1%	+100%
Abandoned/incomplete utterances (per clause)	0.3%	0.6%	+100%
Ellipsis (per clause)	1.9%	3.0%	+57%
Unfilled pauses (per clause)	0.4%	7.8%	+1,850%
Filled pauses (per clause)	0.7%	6.2%	+786%
Repetitions (per clause)	0	2.2%	n/a
Reformulations (per clause)	0	2.8%	n/a

The results indicate that the texts produced after the training were on average 24% longer and contained considerably more instances of repetitions, reformulations, filled and unfilled pauses and interjections. Moreover, the texts contained more spoken discourse markers, first-person, emotive and interrogative utterances, as well as more instances of spoken grammar such as ellipsis, abandoned or incomplete utterances, and simple linking words. At the same time, there were fewer passive verb forms, as well as embedded and subordinate clauses, which are more characteristic of written language.

It was further found that those texts which the item reviewers had given a score of '0' were the shortest (M=211 words), had the most instances of embedded clauses and the fewest instances of emotive or interrogative utterances; they were largely narrated in third person (50%, compared to 24% for texts scoring '1' and 19% for texts scoring '2'). They also contained few discourse markers

and no interjections, incomplete utterances, unfilled pauses, repetitions or reformulations. Furthermore, the six texts that scored '0' did not reflect the genres stated in the specifications. For example, two texts were monologues, but the item writers defined them as extracts from informal conversations. However, an extended monologue embedded in an informal conversation is implausible and did not reflect any of the specified genres. For the other four texts, the purpose and the situation of speaking were difficult to identify from the text or item instructions, thus the genre was unclear.

The difference between texts that scored '1' and '2' on the text authenticity criterion was more subtle. Before the training, band '1' texts generally contained more instances of spoken language than band '0' texts but fewer than band '2' texts; they also had more written language features compared to band '2' texts such as subordinate clauses (28% compared to 24% for band '2' texts) and passive verb forms (0.8% compared to 0.4% for band '2' texts). Texts that were awarded band '2' before the training were generally the longest (M=251) and contained the largest number of instances of spoken language features with fewest instances of written language.

After the training, band '1' texts had more instances of spoken language features compared to band '1' texts produced before the training. Moreover, the number of instances was higher in some categories than for band '2' texts. For example, post-course band '1' texts contained more exclamations, questions, and unfilled pauses. At the same time, band '2' texts produced after the training had substantially more simple linking words (3.6% compared to 2.1% for band '1' texts), and first-person utterances (57% compared to 40% for band '1' texts). Furthermore, post-training band '2' texts had twice as many spoken discourse markers, interjections, incomplete sentences, repetitions, and reformulations, compared to post-training band '1' texts. These band '2' texts also contained a wider range of different spoken features, compared to band '1' texts: most comprised both filled and unfilled pauses, repetitions, reformulations, and a range of discourse marker types, while the band '1' texts generally contained a narrower range of features overall and within each text.

Qualitative data analyses

We present the qualitative findings by the trainee profiles (Table 2), with three types of data discussed in chronological order, i.e. findings from the (1) pre-course interviews; (2) online group discussions held during the 'Testing Listening' module of the course and concerned with listening text authenticity; and (3) post-course interviews.

Profile a: Low pre-course text authenticity scores, high post-course text authenticity scores

In the pre-course interviews, four of these trainees did not discuss listening text authenticity, while James and Henry acknowledged its importance but said that producing authentic-sounding texts was difficult. They mostly dwelt on what the listening texts should *not* be like – for example Henry said *"…what will be difficult … is keeping it authentic, it can't be too stilted, it can't be as though somebody were reading it"* – but were unable to elaborate on what authentic-sounding texts *should* be like.

During the course, in small discussion groups, participants provided peer feedback on the listening texts they had initially developed for the pre-course assignment and then improved following the course input on listening text authenticity. The 'profile a' trainees discussed a wide range of spoken language features which had been introduced in the lecture and which they subsequently employed to make their listening texts sound more authentic: pauses, fillers, hesitations, false starts, repetitions, and redundancy. They also talked about grammar features such as short and simple sentences. In terms of vocabulary, they discussed the use of colloquial language and language typical of the genre. At times, these trainees demonstrated an understanding of spoken language features richer than the course input. For example, Henry realised the importance of the location of pauses in the text to make it authentic-sounding*: "Mine* [my pauses] *would benefit by paying attention to where these would appear in natural speech",* while Ted reflected on differences in the order of ideas in spoken and written texts:

> *I've tried to make it more natural by playing with the order of ideas within a sentence, like the speaker doesn't quite get everything in the right order first time.*

In the online discussions, participants demonstrated an understanding of the role of genre in determining listening text features. For example, they discussed the difference between scripted and unscripted listening texts: "*That looks fairly authentic to me given the context, i.e. it's a scripted advert"* (Matthew). At the same time, the trainees expressed the need for more examples of transcribed spoken texts to get a better sense of spoken language features for each genre, with the concept of redundancy being particularly difficult to understand: *"Would be nice to have some examples of redundancies … I don't naturally think of people being redundant in their speech"* (Adam).

In their post-course interviews, these trainees also reported employing a whole range of grammatical, lexical and discoursal spoken language features, for example:

> *… the sentences are not really full sentences, there's a lot of redundancy like 'yeah, like I said', 'yeah I dunno', 'bit of a shame really', these little phrases that English people throw about* (Lucas)

Similar to what was found for their discussions during the training, these trainees adopted a deep, reflective approach to the authenticity requirement. For example, they were aware that the texts will need to be studio-recorded, and incorporated guidance for voice actors in their scripts. Emily also extended the authenticity concern to the items by asking herself: *"If I really wanted to know about this piece of technology what would I want to know about it? What would I be listening for when in a real situation?"*

Earlier, we described different techniques for listening text sourcing and production: use of *genuine* texts, genuine text *adaptation*, *improvising, semi-scripting,* and *scripting* (including text *authentication)*. The post-course interviews revealed that none of the course participants used or adapted a genuine text. In the 'profile a' group, Emily, however, reported a hybrid of adaptation and semi-scripting: she used a genuine YouTube video clip to generate content points, and then recorded herself speaking from those content points. Three 'profile a' trainees reported using the semi-scripting technique, while Lucas used a variation on the semi-scripting technique, whereby he vocalised his text from content points, but without recording it. Only one trainee in this group, Ted, authenticated a written text. Although some of these trainees found producing an authentic-sounding text difficult, they were convinced that the effort was worthwhile*: "You need to get the text right, and it has to sound natural … you have to make it sound right"* (Henry).

The post-course interviews of these trainees were also analysed for course features which they felt had particularly influenced their item-writing skill development. The trainees indicated that the course input on spoken language features was valuable for producing authentic-sounding listening texts. In addition, extensive item-writing practice was described as possibly the most rewarding course feature because "*the more we do, the more comfortable we will feel with them* [writing items]" (Henry). The item-writing practice during the course included a collaborative element whereby trainees shared drafts of their items for peer discussion and feedback. Most of those participants who managed to improve their text authenticity score described this group work as beneficial: "*sometimes people point things out and help you out and I think that can be quite helpful as well*" (Adam). Adam furthermore said that seeing other participants' items helped him gain confidence in his item writing as he could see some "*pretty poor*" items and realized "*it* [producing poor items] *happens to everyone*". Finally, the data indicated that the tutors' feedback played a particularly major role in developing trainees' ability to produce authentic-sounding texts.

Participants received tutor feedback on their pre-course texts as well as the texts produced during the listening module. Most mentioned that the feedback was beneficial, for example Lucas said on this topic: "…*to know what you did wrong so then the next time I do it, I know what to pay attention to"*.

Profile b: Low text authenticity scores pre- and post-course

In the pre-training interviews, two 'profile b' trainees did not mention text authenticity, while seven admitted their texts did not sound authentic, for various reasons. For example, Arthur prioritised other specification requirements which he found incompatible with text authenticity, Olivia did not know how to make a text sound authentic, while Nathan seemed to underestimate the importance of the authenticity requirement: *"I thought this is a listening task, not supposed to be completely natural sounding really*". Two trainees expressed regretting not having tried to comply with the authenticity requirement:

> *… from a position of authenticity, I really should have tried to get into some kind of mental space that I tried to make it* [the text] *as authentic as possible. I didn't, actually* (Mason)

During the course, these trainees were active in providing each other with feedback on listening text authenticity, and they discussed a wide range of spoken language features introduced during the training. For example, Daniel commented on his revised text:

> *The changes are in bold and include: corrections, afterthoughts, simple conjunctions, emphatic language, hesitations and asides. There is also language typical of the context like 'indeed'.*

However, these trainees' discussions were somewhat different from the 'profile a' trainees': they did not seem to attach as much importance to the text genre, and their understanding of spoken language features did not extend beyond the course input. These trainees seemed to follow text authenticity recommendations more mechanically by incorporating as many spoken language features as they possibly could, irrespective of their appropriacy.

After the training, most 'profile b' trainees used the authentication technique to produce listening texts, as they reported in their interviews. Two trainees mentioned awareness of other techniques but decided against them: *"I must admit I didn't use the suggested method of recording it* [the text] *first, I felt I could write something reasonably authentic-sounding without doing that"* (Jake). When using a technique other than authentication, 'profile b' trainees reported difficulties with it. For example, Stanley attempted text improvisation but, after transcribing the recording,

discovered that it was not possible to generate items on the basis of the text. Two trainees reported attempting semi-scripting, but then abandoning the technique: "*It didn't … work, I got a little bit self-conscious*" (Mason).

Many trainees in this group stated that producing listening texts was one of the most challenging aspects of item writing. However, the challenge did not lie in producing texts per se, but rather in producing authentic-sounding texts that would at the same time meet the specification requirements (e.g. word frequency characteristics of both the text and items). One difficulty reported by these trainees was making sure that the text's lexis complied with the word frequency requirements, something that many participants found to be incompatible with the idea of text authenticity. The trainees also reported difficulties with item generation, for example, how to include enough information points to be tested in items while keeping a natural text flow. The key challenge, however, lay in including the required number of pieces of distracting information without making the text sound contrived. Many trainees felt that distractors were not a feature of genuine texts and that by including distractors, they were compromising text authenticity:

> *When I was writing this* [the listening text] *I wrote it with a lot of other information that I feel was not necessarily the way that somebody would speak naturally, like I included things that I thought were unnatural so that the candidate would have an opportunity to select a statement that was accurate* (Arthur).

Profile c: High text authenticity scores pre- and post-course

Of the two trainees in this profile group, only Logan volunteered to be interviewed both before and after the training. Pre-course, Logan's discussion of the listening text's authenticity was markedly different from that of the trainees in the other two groups:

> *[I] approached it … trying to remember where I've heard things on the radio … a monologue instruction on the radio, a monologue … so I thought of something going down that line… basically, I walked around the room, so having this conversation in my head on how it would actually sound and then I was sort of able to complete it as well… I was writing it at the same time as writing notes on how to do it … doing the speaking myself, sounding like a crazy person was I think possibly a necessary process as well, how does this actually sound rather than how does this look on paper… I actually did enjoy that. I almost felt like I was doing the writing stuff for the radio or something, it felt very different and that's what I'm enjoying about this.*

Logan was convinced that text authenticity is an important aspect of listening assessment and he enjoyed the process of creating an authentic-sounding text. He was aware of the text specification requirements ("*monologue*", "*on the radio*"), and he drew on his familiarity with real-life radio broadcasts to produce an authentic-sounding text. Moreover, Logan intuitively arrived at an approximation of a semi-scripted approach to listening text production: he wrote notes on the text content and then sounded the text out to make sure it sounded authentic. Notably, Logan was the only interviewee who elaborated on his approach to listening text development before the training.

Logan was very active in discussing text authenticity during the course, revealing awareness of a wide range of spoken language features as well as attention to text genres. He provided detailed peer feedback, demonstrating a deep reflective approach to authenticity, for example: *"…less repetition would be better and more natural"; "[p]unctuation I feel is necessary for the person /actor reading the text to emphasise what you were trying to achieve - in meaning and tone".* This is similar to 'profile a' trainees, who also demonstrated deeper understanding of text authenticity, and different from 'profile b' trainees whose understanding seemed more mechanical.

In the post-course interview, Logan reported using the same technique as he did pre-course – vocalising the text from content points without recording it. Even though he said producing an authentic-sounding text was difficult, he felt the authenticity requirement was important, so he paid particular attention to it.

Discussion

This study investigated the possibility and effectiveness of training item writers to produce authentic-sounding texts for testing listening – a previously unexplored topic. It was found that the training had a statistically significant positive effect on the perceived authenticity of the produced listening texts, with medium-to-large effect size. Analysis of the listening texts' linguistic features provided support for the statistical findings, demonstrating that the texts produced after the training contained, on average, more instances of spoken language features and fewer instances of features characteristic of written language. The fact that texts produced post-training were substantially longer might also point to them sounding more authentic; in Gilmore's (2004) study, for example, genuine spoken texts were twice as long as their textbook equivalents because the genuine texts were less straightforward and contained hesitations, reformulations, and other features of spoken discourse.

More detailed analysis revealed that texts which scored '0' on the authenticity criterion did not reflect any of the genres stipulated in the specifications and contained the lowest numbers of spoken language features. It should be noted that no text scored '0' on the authenticity criterion after the training, which suggests a positive influence of the training on participants' ability to produce authentic-sounding texts.

Before the training, the texts that scored '1' generally contained fewer instances of spoken language features and more instances of written language features compared to the pre-course texts that scored '2'. After the training, however, the number of instances for some of the spoken language features was higher in band '1' texts than in band '2' texts. A potential explanation for these texts' band '1' score on the authenticity criterion is that the spoken language features were incorporated less skilfully than in band '2' texts. For instance, Carter & McCarthy (2007) found that unfilled pauses normally occur on clause boundaries in genuine spoken discourse; in the present study, only half of the pauses occurred on clause boundaries in the post-training band '1' texts. In band '2' texts, 80% of the pauses did. The same is true for repetitions: Carter & McCarthy (2007) found that repetitions in genuine spoken texts most often occur at clause beginnings; in the present study, 58% of all repetitions in the band '2' post-course texts occurred at clause beginnings, while only 28% of repetitions in the band '1' texts did. As for exclamations and questions, which were more abundant in band '1' texts, their large number could have been perceived by the reviewers as excessive and, thus, artificial. Moreover, band '1' texts that were produced after the training generally contained a narrower range of spoken language features than band '2' texts, with some features over-represented and others under-represented (e.g. spoken discourse markers and reformulations).

Overall, it seems that it is not the abundance of spoken language features or the presence of one particular feature that makes a listening text sound more authentic but the appropriate and skilful use of a wide range of such features. The item writers whose texts scored '2' on the authenticity criterion after the training might have developed a deeper understanding of what makes a spoken monologue sound authentic, which allowed them to introduce spoken language features in their texts in a more natural manner. The interview and online discussion data also supported this: those trainees who optimally developed their ability to produce authentic-sounding listening texts (scoring '2') revealed more nuanced understanding of spoken language features, while the trainees whose ability to produce authentic-sounding texts had scope for further improvement (scoring '1') demonstrated more mechanical approaches to ensuring text authenticity.

Before the training, participants were largely unable to elaborate on what an authentic-sounding text should be like, confirming Gilmore's (2015) observation that "native speaker intuitions about language and speech behaviour are notoriously unreliable" (p.515). Indeed, it is often taken for granted that item writers, who are either native or highly-proficient speakers, are familiar with spoken language features through their spoken communication experience. However, proficient speakers normally concentrate on the meaning rather than the form of oral messages. They 'edit out' features of spoken language when communicating in real life and are not necessarily consciously aware of them. The pre-course interview findings, therefore, confirmed the need for explicit training of item writers on spoken language features and provided support for the inclusion of information about phonological, grammatical, lexical, and discursive features of oral texts in the listening module of the course. The online group discussions and post-course interviews suggested that training in spoken language features had a positive effect on developing item writers' ability to produce authentic-sounding listening texts. A wide range of features was discussed by the participants, and trainees whose texts gained higher scores on text authenticity demonstrated a better understanding of spoken language features as well as higher awareness of their role in the text production process, including the need to provide guidance for voice actors.

The post-course interviews revealed that none of the participants used or adapted a genuine sound file for their listening task. The texts of participants who used the semi-scripting technique (Buck, 2001) all gained '2' on the authenticity criterion, which might suggest that the semi-scripting technique brings excellent results in producing authentic-sounding listening texts. At the same time, the technique was not attempted by many participants and did not work for everyone who attempted it. It seems that item writers might require stronger encouragement in producing semi-scripted texts, and also more extensive training in using the technique. A variation of the semi-scripting technique, whereby texts are vocalised without being recorded, worked for several participants and might be considered a viable alternative. The only participant who used the improvisation technique did not get a high text authenticity score because the improvisation did not result in enough testable content. Written text authentication, although most popular, did not bring good results in all but one case, which aligns with Wagner's (2018) observation that written text authentication cannot satisfactorily reproduce discursive and organisational features of listening texts. It should be noted, however, that the present study did not experimentally control for text production technique, and thus these results should be treated as indicative rather than prescriptive, and more research into listening text production techniques is needed to confirm or disprove our findings.

After the training, participants reported that the need to reconcile text authenticity with other specification requirements makes producing authentic listening texts challenging. This might suggest that the problem may not (just) stem from insufficient item writer skills but could (also) relate to specifications or characteristics specific to the act of testing. For instance, participants particularly struggled with the requirement for the text to have distractor-related information – typically needed for selected-response item formats. The problem of distractors is also noted in the literature (e.g., Field, 2019a); distractors are not an authentic feature of most real-life texts, so test designers need to carefully balance the requirement of distractor information in listening texts if they want the texts to have authentic characteristics. Overall, the findings regarding the challenges of reconciling text authenticity with item creation empirically demonstrate the tension between listening text authenticity and the scope of a text to enable high-quality items, as previously speculated in the literature. However, we would also like to emphasize that the above problems were generally reported by participants whose listening texts gained band '1' post-course on authenticity, while high achievers seemed to have developed some coping strategies to reconcile text authenticity and other specification requirements.

The post-training interviews indicated that the course input on spoken language features, as well as the listening text-writing activities with peer- and tutor-feedback were course features that enabled the participants to produce authentic-sounding texts following the training. Tutor feedback played a particularly important role, with most participants whose texts scored '2' after the training reporting to have attended to the feedback.

Conclusion

Although Salisbury (2005) discussed "an ear for 'speakerly text'" (p.293) as a pre-existing item writer characteristic necessary for producing listening tasks, findings from this study suggest that, for a considerable number of item writers, their ears can be attuned through relevant training, and that the ability to produce (more) authentic-sounding texts can be developed through practice infused with effective item-writing techniques and reinforced by informed feedback. Given the high demand for good item writers (Buck, 2009), it seems important to be able to increase the item writer pool by not only recruiting people with 'pre-existing item-writing abilities' – which would have meant very few suitable individuals in this study as only two of the 25 recruited trainees produced satisfactory texts before training – but also by developing the item-writing skills of generally suitably-qualified people through targeted training.

At the same time, our study revealed that the extent of item-writing skill development was unequal among participants: while 1/3 of trainees seemed to have competently developed their ability to produce authentic-sounding texts following the course, 2/3 still had scope for further improvement. In particular, twelve participants produced texts that scored '1' on the authenticity criterion both before and after the training. It should be noted, however, that this does not mean these participants did not develop their item-writing skills at all. In fact, analysis of the texts produced by these participants demonstrated an increase in spoken language features following the training. In addition, the analysis of these participants' group discussions and interviews revealed a deeper understanding of what text authenticity entails after the course. It might be that these participants need more time for the effect of the training to translate into a 'fully-developed' ability to produce authentic-sounding texts. This is not entirely unexpected as several experts have recognised the need for lengthy item-writing and revision practice to reach item-writing mastery (e.g., Spaan, 2007), whereas the course was restricted in time. Ultimately, longitudinal research would be required to determine whether the latter participants would be able to fully master the skill at a later stage. Another limitation of this study is the relatively small sample size (25 participants), which requires caution in generalizing. This sample size, however, resulted from practical considerations as it is difficult to manage a moderated online training course with large groups.

Our findings have important implications for the theory, as well as the practice of listening assessment. They provide counterevidence for the exclusive use of genuine texts under the motto of validity and authenticity. While genuine texts certainly have their place in assessing listening, they might not always be suitable for item generation, as previously argued in the literature (Buck, 2018; Richards, 2007; Widdowson, 2003), since they might become inauthentic when taken out of their original context and might not allow for adequate language sampling to generalise test results to the TLU domain. Our study indicates that purpose-developed listening texts can be created in such a manner that they are perceived as authentic, and that item writers can be trained to produce such authentic-sounding texts for listening assessment. Findings from this study also indicate that item specifications which, on one hand, demand text authenticity but, on the other hand, preclude it with overly restrictive requirements regarding issues such as vocabulary frequency or in-text distractors, might contribute to the challenge of text authenticity in listening assessment. While specifications help ensure construct-relevant and consistent items, test form comparability, etc., a balance needs to be found in how constraining they are. From a practical point of view, this study offers many insights into the pre-/post-course performance of those trainees who optimally developed their ability to produce authentic-sounding texts and of those who did not, including their understanding of spoken

language features, their item-writing approaches, and their challenges in developing listening texts. These insights will be useful to large-scale testing organisations and educational institutions who produce their own listening tests and are planning to conduct item-writer training or looking to improve their existing item-writer training practices.

Funding

The authors acknowledge the role of the British Council in making this study possible. The British Council provided a research grant which enabled the first author to conduct part of the study under the Assessment Research Awards and Grants programme 2018. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the British Council, its related bodies or its partners.

References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, *3*(1), 3–30. https://doi.org/10.1207/s15434311laq0301_2

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303–310.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension* [Technical Report]. University of Maryland Center for Advanced Study of Language.

Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox v. 4.x [software]. http://corpora.lancs.ac.uk/lancsbox

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus (TOEFL Monograph Series No. 25). ETS.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Buck, G. (2009). Challenges and constraints in language test development. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp.166-184). Multilingual Matters.

Buck, G. (2018). Preface. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp.xi–xvi). John Benjamins.

Buendgens-Kosten, J. (2014). Authenticity. *ELT Journal*, *68*(4), 457–459. https://doi.org/:10.1093/elt/ccu034

Carter, R.A., & Mc.Carthy, M.J. (1997). *Exploring spoken English*. Cambridge University Press.

Carter, R.A., & McCarthy, M.J. (2007). *Cambridge grammar of English: Spoken and written English grammar and usage.* Cambridge University Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clark, M. (2014). The use of semi-scripted speech in a listening placement test for university students. *Papers in language testing and assessment*, *3*(2), 1–26.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp.77–151). Cambridge University Press.

Field, J. (2019a). *Rethinking the second language listening test: From theory to practice*. Equinox.

Field, J. (2019b). *Authenticity: The elephant in the language tester's room.* Paper presented at the CRELLA Spring seminar, Luton (UK).

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal, 58*(4), 363-374.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, *40*(2), 97–118. https://doi.org/10.1017/S0261444807004144

Gilmore, A. (2015). Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials. *Language Teaching, 48*(4), 506-530. https://doi.org/10.1017/S0261444815000269

Green, A. (2014). Adapting or developing source materials for listening and reading tests. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.830–846). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla087

Green, R. (2017). *Designing listening tests*. Palgrave Macmillan.

Lynch, T. (2009). *Teaching second language listening*. Oxford University Press.

MacDonald, M., Badger, R., & White, G. (2000). The real thing? Authenticity and academic listening. *English for Specific Purposes*, *19*, 253–267. https://doi.org/10.1016/S0889-4906(98)00028-3

Morrow, K. (1977). Authentic texts and ESP. In S. Holden (Ed.), *English for specific purposes* (pp.13–15). Modern English Publications Ltd.

O'Neill, L.D., Mortensen,S.M.R., Nørgård, C., Øvrehus, A.L.H., & Friis, U.G. (2019). Screening for technical flaws in multiple-choice items. A generalizability study. *Dansk Universitetspædagogisk Tidsskrift, 26*, 51-65.

Pinner, R. (2014). The authenticity continuum: Empowering international voices. *ELTED Journal*, *16*, 9–17.

Richards, J. C. (2007). Material development and research: towards a form-focused perspective. In S. Fotos & H. Nassaji (Eds.), *Form-focused instruction and teacher education: Studies in honour of Rod Ellis* (pp.147–160). Oxford University Press.

Salisbury, K. (2005). *The edge of expertise? Towards an understanding of listening test item writing as professional practice* (Unpublished doctoral dissertation). King's College, University of London.

Shohamy, E., & Reves, T. (1985). Authentic language tests: where from and where to? *Language Testing*, *2*(1), 48-59. https://doi.org/10.1177/026553228500200106

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly, 4*(3), 279-293. https://doi.org/10.1080/15434300701462937

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, *2*(1), 31–40. https://doi.org/10.1177/026553228500200104

Thorn, S. (2018). A new matrix for testing listening using authentic recordings. *TEASIG Newsletter, 64*, 20-23.

Wagner, E. (2014). Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal*, *5*(2), 288–311. https://doi.org/10.1002/tesj.120

Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp.103–123). Bloomsbury.

Wagner, E. (2018). A comparison of L2 listening performance on tests with scripted or authenticated spoken texts. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp.29–44). John Benjamins.

Wagner, E., & Ockey, G. J. (2018). An overview of the use of authentic, real-world spoken texts on L2 listening tests. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp.13-29). John Benjamins.

Wagner, E., & Toth, P. D. (2014). Teaching and testing L2 Spanish listening using scripted vs. unscripted texts. *Foreign Language Annals*, *47*(3), 404–422. https://doi.org/10.1111/flan.12091

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

White, G. (2018). Authenticity in listening assessment. In J. I. Liontas (ed.), *The TESOL encyclopedia of English language teaching* (Vol. 34, pp.1–6). https://doi.org/10.1002/9781118784235.eelt0616

Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford University Press.

Widdowson, H. G. (2003). *Defining issues in English language teaching.* Oxford University Press.

Appendix 1: Listening task specifications

Listening comprehension task for a general English proficiency test, adult candidates, unspecified nationality

Task description	Gap-fill
Skill focus	Ability to locate and record specific information from a text
Task level	B1
More information about the task	Candidates have a set of notes or sentences, summarising the key content of the text, from which six pieces of information have been removed. As they listen, they fill in the numbered gaps with words from the text which complete the missing information. This may be key pieces of information about places and events, or people talking about courses, trips, holiday activities or other types of factual information. The words candidates need to complete the gaps are heard on the recording: single words, numbers or very short noun phrases.
Instructions to candidates	You will hear … (specify the speakers and the situation., e.g. *a woman talking on the radio about a new sports centre*). For each question, fill in the missing information in the numbered space with a maximum of 3 words or a number.
Listening input specifications	

Text type	A monologue
Text length	max. 300 words
Lexical level	K1 to K3
Grammatical level	A1 to B1
Topic	From the list of topics for B1 level
Text genre	A monologue: lectures/presentations, TV/radio programmes, short talks.
Text authenticity	The text should sound like authentic spoken English (according to the genre) and not a written script read out.
Function	From the list of functions for B1 level
Item specifications	
Item type	Gap-fill, each gap to be filled with a maximum of 3 words or a number heard in the text. The items are either a set of notes or sentences. Items should follow the order of the text.
Distractors	Distractors will be used in the input text. Each item (except for proper names that are spelt out) should have 1 or 2 distractors.
Items per task	6 in total
Stem length	Maximum 10 words including the key; the stem should not literally repeat what is heard in the text but should be a paraphrase
Stem lexical level	K1 to K2
Stem grammatical level	A1-A2
Response type	Concrete information
Response length	Maximum 3 words or a number from the text
Response lexical level	K1 – K2 (except for proper names that are spelt out, there should be no more than 1 item of this kind per task).