When the test developer does not speak the target language: The use of language informants in the test development process

Abstract

It is not unusual for tests in less-commonly taught languages (LCTLs) to be developed by an experienced item writer with no proficiency in the language being tested, in collaboration with a language informant who is a speaker of the target language but lacks language assessment expertise. How this approach to item writing works in practice, and what factors play a role in it, is largely unrecorded, as are item writing processes and practices in language assessment in general.

Through a case study approach, this study sought to gain insights into test development practices in cases when essential item writer traits are spread across different people. Seven in-depth interviews with language assessment specialists and language informants involved in LCTL reading test development revealed a number of specific characteristics, and also challenges, to test developer recruitment and test development in this context. Findings indicate that this inherently collaborative approach brings with it a sophisticated system of "checks and balances" which may benefit item writing in some respects.

Background

The language testing profession advocates that individuals involved in test development be held to high professional standards, which means that qualified item writers are often sought after (Buck, 2009). According to the industry's standards of practice, supported by empirical research (Kim, Chi, Huensch, Jun, Li, & Roullion, 2010), it is recommended that test writers have two major competencies: language teaching experience “at the level the assessment is aimed at” (EALTA, 2006, p. 3), and knowledge of “current language testing theory and practice” (ILTA, 2007, p. 3). In other words, “general proficiency language testers are likely to possess linguistic knowledge, language skills (fluency in the target language, or *access to it through a colleague*), measurement, and research design skills” [emphasis added] (Davies, 1997, p. 82).

Even though language tests are typically developed by native speakers (NSs) or highly proficient nonnative speakers (NNSs) of the target language, such is not always the case. The phrase “*access to it through a colleague*” in the above quote hints at the fact that language testers are not always fluent in the target language. Indeed, a number of studies have looked into the involvement of NNSs in the language test cycle, thereby primarily focusing on NNSs as raters of language tests (e.g., Gui, 2012; Zhang & Elder, 2010), and assuming that the NNS is proficient in the language being tested, which in most studies is English. There is a paucity of research about tests written by NNSs with more limited proficiency in the target language, which is sometimes the case in classroom foreign language teaching contexts (cf. Richards, Conway, Roskvist, & Harvey, 2012), and there is even scanter scholarship on language tests developed by professionals with no proficiency in the language being tested. The latter case is not unusual, however, for tests in less-commonly taught languages (LCTLs), which focus on all world languages except for English and the commonly taught languages of German, French and Spanish (NCOLCTL, n.d.)[1]. Indeed, the ever-increasing demand for language tests is not often met by a workforce that is at once proficient in the target language, locally accessible, and adequately trained in test development. Amongst the extremely scarce publications on such situations, Kennedy and Stansfield (2010), along with Brooks and Mackey (2009), report on assessments of receptive skills developed by

[1] The definition provided here is that of the National Council of Less Commonly Taught Languages. We appreciate, however, that the taught status of languages may vary in time and region, and thus the list may not be fixed.

NNSs with little or no knowledge of the target language.[2] Both papers describe a specific context: proficiency tests in LCTLs developed for the United States Government based on the Interagency Language Roundtable (ILR) scale. The examinees for whom these tests are developed are native (or highly proficient) speakers of English. Kennedy and Stansfield (2010), for instance, outline the development and piloting of a Reading Proficiency Interview for Marshallese, an official language of the Republic of the Marshall Islands in the Pacific Ocean. Other examples, described in Brooks and Mackey (2009), constitute versions of the Defense Language Proficiency Test 5, which is a battery of foreign language tests developed by the US Defense Language Institute, used to assess the reading and listening skills of US Department of Defense linguists. Target languages include – to name only two – Yoruba (principally spoken in Nigeria and Benin), or Cebuano (The Philippines) (http://www.dliflc.edu/dlptguides.html).

For the development of some of these tests, experienced language test developers are paired up with so-called 'language informants'. The former is an assessment literate person in the narrow sense of the word, i.e. someone who possesses "the knowledge and skills required for performing assessment-related actions" such as language test production (Inbar-Lourie, 2012, p. 2923), but who has no or limited knowledge of the target language. The language informant, on the other hand, is a speaker of the target language, but does not usually have a language assessment background (Brooks & Mackey, 2009; Kennedy & Stansfield, 2010). This 'consultant' (the language informant) also has to be proficient in a language that can be used for communication with the language assessment specialist. Together, the language assessment specialist and informant work hand-in-hand to create a foreign language test (Brooks & Mackey, 2009). In other words, while there is sufficient in-house assessment expertise, what needs to be contracted out is knowledge of the language to be assessed. The 'ideal' test writer profile is thus spread across more than one person and together they need to fulfill the job of the test writer. It should be noted,

[2] The term 'NNS' may in fact not be a good fit, in the sense that the person may simply not speak (or know) the language at all.

however, that this is not unique to LCTL testing, but for example also occurs in Language for Specific Purposes (LSP) test development, where language testers may collaborate with content specialists.

However, given scant scholarly attention, it remains unclear what the exact nature is of such collaboration, or what considerations play a role in the collaboration. In addition, the collaboration may bring about some differences with the characteristics of the ‘conventional’ test development process. Therefore, the major aim of this paper is to examine this collaborative approach to test development in order to gain insights into how professional standards of practice are met in such contexts. The research question that guides the present study is: what idiosyncratic issues arise during the test development process when the assessment experts have no proficiency in the target language and the consulting target language speakers are not assessment literate? To obtain a rich and detailed understanding of this unique context, a case study approach (Yin, 2012) was adopted. It is hoped that this paper will provide test development professionals and researchers with an insightful discussion about this method of test development.

Methodology

Case Study Context

As explained by Yin (2012), a case study method is particularly suitable for in-depth empirical research aiming to describe and understand the intricacies of a particular situation, which seemed relevant in this situation given the relative lack of research on LCTL test development.

The setting of the present paper is that of a US-based private testing company, which has been specializing in language proficiency assessment and computer-adaptive testing for more than a decade, having developed tests in over 40 languages. The case study is set within the boundaries of this language testing company, with its particular set of expectations, policies and practices when it comes to LCTL testing. The specific context of the study was the development of reading items in LCTLs, which the company had been contracted to produce by several government agencies. More precisely, the study focused on the development of reading items targeting the lower and mid-levels of the ILR scale, i.e. ILR 0+ "memorized proficiency" to ILR 3 "general professional proficiency".[3]

The target population of the tests is government personnel who are native speakers of American English. The contexts in which these tests are used depend on the commissioning agency. For instance, some test-takers have their reading skills assessed at the end of their language learning training, while others are required to take a reading test on a yearly basis. Similarly, some of these tests are considered high-stakes when their results are used, for example, to inform decisions regarding pay grade and professional assignments, whereas other tests are considered low-stakes.

The format of the item sets is that of a passage in the target language (the LCTL), with 1-3 corresponding comprehension questions formulated in English per passage. The passages are taken from real-life sources such as advertisements, news articles, etc. and are up to 400 words in length. The item types constitute a combination of automatically-scored selected-response and human-scored constructed-response questions – namely, 4-option multiple-choice items and short-answer questions, requiring the answer to be formulated in English and targeting reading for implicitly- or explicitly-stated main ideas and details.

Participants

The case study participants (Table 1) comprised a convenience sample of seven people who were involved in the development of the previously described LCTL reading items. Four of these (two males, two females) were regular employees of the company working as assessment experts. They were approached for the study based on previous professional contact with one of the researchers. They had

[3] Due to the nature of our participants' assignments (see below), the present study specifically covers item development and does not report on the broader test development and quality assurance processes in place at the company and for the reading tests as a whole.

two to five years of experience (M=4) working on LCTL tests as well as on tests in more widely spoken languages. Two of them were mostly tasked with collecting test materials such as reading passages, while the other two primarily did item writing. All were native or highly proficient speakers of American English; however, in this context of LCTL testing, they were NNS test developers in the sense that they did not speak the target language of most of the tests they worked on at the company. These four participants held degrees in TESOL, linguistics, and a foreign language. In the remainder of this article, these participants will be referred to as the test development experts.

The other three participants had been contracted by the company as ‘language informants’, i.e. speakers of one of the languages in which a LCTL test needed to be developed. They were recruited for research participation via the test development experts, i.e. the researchers asked the test development experts to identify a language informant they had worked with. They were foreign-born residents of the United States, and their native languages were Persian Farsi (spoken in Iran), Somali (spoken in Somalia), and Telugu and Hindi (spoken in India). They were also proficient in English, and had lived in the US for periods ranging from three to eleven years. They held university degrees in the social sciences and humanities and their primary occupation was outside of language proficiency testing (as graduate students, teachers, translators).

It was hoped that selection of both test development experts and language informants as participants would offer multiple perspectives on the topic under scrutiny. Despite the limited sample, which reflected the restricted number of professionals involved in this idiosyncratic test development situation and available as volunteers for the study, the in-depth quality of the interviews provided richly descriptive information about the process.

[TABLE 1 HERE]

Data Collection

Semi-structured interviews were conducted by one of the researchers with each of the participants, using Skype because of the dispersion of the participants throughout the US. Two interview guides were prepared in advance (one for the test development experts and one for the language informants), drawing on a review of the test development literature and on feedback gathered during a prior unstructured interview via Skype with one of the test development experts, in which he was asked to describe his assignment. Interviews were first conducted with the test development experts in order to get a detailed account of the item development process as well as information regarding the tasks expected of language informants versus test development experts. This feedback was then used to refine the interview protocol for the language informants, who were subsequently interviewed. The general themes for discussion with the test development experts were: preparation for the item development work, the item development process, language informants' profiles and implications for item development, LCTL item development challenges, identifying reading test input materials, the ILR scale, and testing reading in LCTLs. The interviews with the language informants covered the following broad topics: project preparation, qualifications, LCTL item development challenges, identifying reading test input materials, the ILR scale, and testing reading in LCTLs. The semi-structured nature of the interviews helped ensure that the interviewer consistently collected data on the predetermined topics from all participants, whilst at the same time leaving room for diversions from the interview guide and elaborations on what seemed to be interesting points brought up by the interviewees (Dörnyei, 2007). The latter seemed particularly useful, given the exploratory nature of the study. The interviews with the test development experts lasted approximately one hour each (M=59.5mins). The interviews with the language informants, which contained fewer questions and covered a more limited number of topics due to the more restricted tasks of the language informants, took approximately half an hour each (M=33mins).

Other sources of information included participant observation and document analysis. However, due to space limitations, the focus of this paper is restricted to the interview data, which formed the main data source. Nevertheless, we can confirm that the other sources triangulate with the interview findings.

Data Analyses

Creswell's (2009) steps for qualitative data analysis were followed. The interviews were transcribed professionally and checked by one of the researchers. To gain an overall impression of the data and form some initial insights, time was spent going through the transcripts. Next, a more detailed analysis was conducted using the software NVivo. A coding frame was generated, consisting of a combination of deductive codes established on the basis of the literature and the interview guide and inductive codes emerging from the data themselves. An initial sample of the interview data was double-coded by a colleague familiar with the research context. Any coding differences were discussed, which led to a somewhat revised and refined coding scheme. Both coders then rated a second sample, based on the consensus reached after the first round. This time, only three out of a total of 115 code allocations (2.6%) required clarification. In total, 10% of the recording time was double-coded. Having reached a high level of coding consistency, one researcher continued to code the entire dataset. An overview of the final coding categories can be found in Table 2. The number of comments related to each coding category has been tallied (see Table 2) simply to give an impression of the salience of each (sub)theme.

Finally, additional checks of the data and data interpretations were sought before the writing-up (George & Bennett, 2005) so as to establish the extent to which the present case study may generalize to other situations. This was, for example, done by gathering feedback on a conference poster presentation from test development experts with similar assignments within other testing organizations, and comparison to other conference talks on the theme (e.g., Clark & Sundstrom-Hebert, 2009). This allowed us to follow the two-step process for generalization of case study research put forth by Yin (2012): first refining our claims based on our results, before implicating other similar contexts.

Results

The interviewees offered a range of insights into the profile, duties and working relationship of those involved in test development (i.e., test development experts and language informants) with direct implications for the item development process. Participants also commented to a lesser extent on issues concerning LCTL item development in the specific context of the ILR Scale. A third theme concerned various elements of the test development process itself, more specifically issues regarding LCTL reading input as well as item development. The number of comments the interviewees made on each of these themes and a breakdown of codes are presented in Table 2. Insights gained from the qualitative analyses of the interviews will be structured according to two of the three main themes: test developers, and test development. The theme ‘proficiency scale’ concerned issues with the ILR scale raised by the test development experts, which did not necessarily pertain to LCTL testing. This theme was eventually left out given its niche nature and relatively low occurrence compared to the two other themes.

[TABLE 2 HERE]

Test Developers

Almost half of the total number of comments (337) concerned the tests developers themselves (i.e., test development experts and language informants), specifically their profile as well as their preparation for their given assignment.

Language informants’ and test development experts’ profile. Even though only one of the test development experts in the study had been directly involved in the language informant hiring process, two other test development experts also shared insights about recruitment decisions. They pointed out that the ideal qualifications of language informants varied depending on their assignment (e.g. collecting passages, reviewing items, proofreading), and that the job requirements may be adjusted in function of the test's target language.

> *Test development expert 1*: We use language informants for some different things. And with different languages we are allowed to lift our standards.

The test development experts explained that some language communities have a strong presence in the country in which the test is developed (in this case: the US) and a large pool of qualified applicants to choose from; while for other languages, especially 'never-taught languages', the immigrant community is small and does not offer a large array of applicants to select from.

> *Test development expert 1*: For example, if you are dealing with French, you are going to find a lot more bilingual speakers with PhDs and other things than when you are dealing with some obscure language.

In cases where it was very difficult to recruit language informants, the test development experts reported that there had been no choice but to work with language informants with limited English proficiency.

> *Test development expert 2*: We don't fault it when we work with someone whose English is lower because we know those resources are hard to find.

However, one test development expert explained that the level of English of language informants is crucial for communication with the test development experts, especially when the tests being developed are at higher levels on the ILR scale.

> *Test development expert 3*: [Language informants] have to be able to speak English fluently so that they can communicate sometimes fairly complex issues and topics to us in English. And this

> is especially important at the higher levels. At the lower levels, native language expertise is much more important than metalinguistic knowledge or knowledge of English.

The need for linguistic and cultural proficiency with American English had to be balanced against the need for recent and prolonged contact with the language and culture of origin. The test development experts emphasized that language informants needed to remain up-to-date with changes in the target language, as well as with related socio-cultural events. In practice, at best, language informants have been living outside of the country of the target language for a limited period of time. If that is not the case, it is hoped that they make a sustained effort to keep up with the target language, in order to avoid issues such as first-language attrition. One language informant, for example, reported doing the following:

> *Language informant 2*: I do go to India. I went last year and I'm planning next year to return. But most of the time, I have some Indian friends here, so I speak to them in Hindi. I also watch Hindi movies, Hindi news and sometimes read some newspapers.

The test development experts also stated that the ideal language informant – despite lacking test development expertise – has experience teaching the target language to the target test audience and insights into language learning processes at different levels of proficiency.

> *Test development expert 4*: We prefer to hire consultants [language informants] who have professional experience in the field of language and linguistics, and ideally teaching, because they have an understanding of what students go through as they're learning at different levels and what that looks like.

In the present study, the three language informants, who rated themselves as highly proficient in English, all credited their teaching experience for providing them with the skills to convey meaning about linguistic and cultural topics to nonnative language learners. The Somali language informant even argued that his teaching and translation experience made him a better target language expert than a "simple"

Somali speaker who would speak the language but would not be able to understand the intricacies of language learning and translation.

> *Language informant 3*: Actually, when you have some teaching and some translating experience on your back, and this is about teaching foreign language, and translation is in play also, you will have an edge as opposed to those [without teaching experience] who might not know what exactly translation means and what it means to be a new language learner.

On the other hand, regarding their own qualifications, the test development experts reported having a background in (applied) linguistics, and experience in test development. Three of the test development experts credited their foreign language learning and teaching experience as helping with their current profession; in other words, having experienced and witnessed second language acquisition provided them with insights into the different levels of the proficiency scale, as well as an awareness of the limitations of an English-centric perspective.

> *Test development expert 1*: You can't expect that everything's the same for all languages. I just think that kind of realization is important.

Additionally, the test development experts felt that prior personal or professional experience with foreigners facilitated communication with language informants, which was corroborated by two of the three language informants.

> *Language informant 3*: All [test development experts] speak English. And I am lucky to have some [test development experts] that their first language is not English; and those guys understand how languages differ and how you render stuff from one language into another, and the difference that it might present in terms of concept and transliteration to our rendering.

Language informants' and test development experts' project preparation. Once the language informants are recruited, they are trained by test development experts on various aspects of test development, according to the task they are assigned to. In this case study, the initial training primarily centered on the proficiency scale, namely the ILR scale. The test development experts reported adapting ILR descriptors to an exam context and elaborating on concrete characteristics that would help language informants select test material. Thus, the training was simplified as much as possible given the low assessment literacy of most language informants.

> *Test development expert 3*: And during the training, we (…) [work] with [language informants] through examples, and (…) we [get] rid of talking about the ILR scale in abstract terms entirely. We also try to avoid linguistic terminology or [company] terminology as much as possible.

The three language informants in this study reported no issues with the training, and overall, felt that the training had been useful for their upcoming assignments.

> *Language informant 1*: You know how I look at [the training] now? It's like you learn different sports by playing it; and that was it—we learned basically the moves from the training.

The test development experts, themselves, prepared in a different manner for a new project. They revealed that they usually begin by researching the language and the culture of their assignment.

> *Test development expert 1*: Ideally we will have enough preparation time to do a little bit of research on the target culture and get a good idea about the structure of the language and the script, and basically have a launching point to start from. And it helps a lot in working with language informants if you have some knowledge of that language.

Indeed, the language informants seemed to value test development experts who demonstrated knowledge and interest in the target language and culture.

> *Language informant 1*: The [test development expert] was already very familiar with a lot of [cultural] differences that exist. (…) she was always very curious to learn these things.

Test Development

The test development experts and language informants also discussed several aspects of test development. Their comments specifically focused on the two major phases of test development, namely: the collection and adaptation of reading input materials, and the item writing phase, which are both guided by test specifications.

Reading input materials. The largest number of comments (113) was made on locating materials. The general process involves language informants being given a list of requirements for passages and adequate texts. Test development experts then use the same checklist to determine the acceptability of the submitted passage by looking at an English rendering of the target language text.

> *Test development expert 1*: So basically our language informants are given different requirements and they are sent out to find authentic texts that we'll use on our tests ... Then they will find those and create renderings and submit them to us, and we'll basically review the renderings and determine whether they're actually at the level that they're supposed to be and whether they've got an adequate amount of assessment points to be used on the test.

The language informants reported finding the overwhelming majority of passages online. This allowed test development experts to check the authenticity of the source.

> *Test development expert 4*: We have the language informants submit the source (…) from where they got the passage. (…) We always take their copy from the website and then (…) compare against the passage that they submitted, just so that we can see if there's any changes.

Several considerations on passages, however, were raised during the interviews, most notably with regards to societal factors influencing the language and availability of input materials (85 comments on the subcode *sociolinguistics*). For instance, all but one participant alluded to the fact that the political situation tied to the target language sometimes affected the availability of passages at certain levels or for certain topics. For instance, one test development expert recalled the political controversy surrounding Uighur (spoken by a Muslim minority in China) and the ensuing difficulties in gathering passages on various topics.

> *Test development expert 1*: In the case of Uighur, a big challenge was getting hold of the materials because people could be prosecuted for writing the sorts of things in Uighur that we might have used on the test. (…) Any language from a place where there's a lot of political problems going on and a lot of war and things like that, it is going to be hard to find passages about geography, or about art and culture that aren't somehow tied into politics and terrorism and things like that.

All participants discussed issues related to the occurrence of other languages in target language passages. For instance, since the test-takers are presumably English native speakers, passages including English cognates or English words and phrases were avoided. Hindi was a good example of a language with a lot of code-switching with English.

> *Test development expert 2*: We had to reject a lot more passages than other languages, and talk to the native speaker and just try to say: "Is there any way we can [find] more passages with not so much English borrowed directly in it?" And they said: "Sure, but it's going to take a lot longer", and it did.

Word borrowings from languages other than English were also sometimes avoided, since these may advantage test-takers who might know those languages (versus those who don't).

> *Language informant 3*: Somalis in the south have a lot of borrowings from Italian; Italians colonized them. And Somalis in the north where Somaliland is now, they have a lot of borrowings from English. Somalis in Djibouti have a lot of borrowings from French. And all of

> them, since they were Muslims before the colonialization, even now they have borrowings from Arabic. (…) The cognates shouldn't be numerous to an extent that they give away the passage.

A further issue was that of language variation. The three language informants reported being watchful regarding dialectical variations when selecting passages. In particular, for languages that were spoken in more than one country, language informants from different regions would cross-check the material. The following test developer explained why it became necessary in Modern Standard Arabic (MSA).

> *Test development expert 1*: So we've got one language informant who submitted that passage and so either they overlooked these expressions that were in a different dialect or they didn't consider it to be a different dialect. And another language informant would look at it and say, "Oh! That's not going to work. That's not MSA." And so, it came down to basically finding another language informant to be the tiebreaker.

In all the aforementioned cases of language (variety) switching, if equivalent words/phrases could be found in the target language, the language informant was asked to edit the passage. However, the risk was that such a decision would lead to language that felt contrived, considering that the authentic passage originally included these instances of language (variety) switching. Eventually, if the switches were too significant, the passage would be discarded.

> *Test development expert 4*: We're careful to ask whether or not that's a typical usage. We don't want to just arbitrarily change it all (…).

> *Language informant 2*: But at times I used to have problems to replace the [English] words because if I replace the word, it could become a higher level, especially like things they don't use in common usage, so it wouldn't be a level-one passage. So then I had to reject [the text] (…).

Furthermore, all the participants made a point of selecting passages that were directed towards an audience in the country where the language was spoken. For example, the Farsi language informant explained that passages targeting the Iranians who live in the Los Angeles area were avoided, since most of the diaspora moved to the US in the 1970s, and the Farsi they speak and write is not always representative of the Farsi currently spoken in Iran.

> *Language informant 1*: In my selection, I was very careful to make sure [the passage] is directed toward people who are inside of Iran. But I never choose, let's say a newspaper or something online that's written by the Iranians who live in Los Angeles because most of them came here back in the '70s, so that's not a good thing to use. (…) Very dated, and in a funny way.

In contrast, the test development experts paid attention to the opposite situation and asked language informants to avoid words or phrases that had recently become popular and that might not stand the test of time.

> *Test development expert 2*: You just have to go back and say, "Is this common? Would a test-taker even know what that meant? Is it trendy?" And sometimes you can find that out, and they'll say, "Yeah, it's just something lately that people are saying." Especially with new languages like Kazakh... We'll often find out that we shouldn't use a word because it's too new.

Another issue touched upon by all participants (16 comments) relates to the writing or script. Two interviewees noted difficulties in finding suitable passages in languages with a recent writing tradition, such as Somali.

> *Language informant 3*: There's no set standard, do you know what I mean? (...) So it's like you can see a complete passage without a comma or a period or a capital letter.

Because of the various issues discussed above, suitable, authentic passages sometimes proved difficult to locate (especially at lower levels). In such cases, the test development expert resorted to purpose writing and asked the language informant to script a text (which two of the three language informants reported having done when passages were difficult to find at lower levels). The test development expert usually refrained from giving an English passage to translate for fear of influencing the outcome and compromising the authenticity of the text. The test development experts highlighted that one of the main issues to watch for when resorting to purpose writing was to keep the text as authentic and culturally plausible as possible. They encouraged the language informants to think of common text types in the target language that could serve as a model, which was not always easy.

> *Language informant 3*: The challenges [of purpose writing] are that the passages of the type that you are required to purpose write were not available at all. (…) You don't have a model because you couldn't find it. If you could find it, you could do it, and scripting wouldn't have been needed.

Finally, all participants discussed issues related to renderings. This code occurred 53 times, and was found to co-occur with both "locating passages" and "item writing", meaning that renderings affect both aspects of test development. In their explanations of what constitutes a rendering, the test development experts contrasted it with a classic translation.

> *Test development expert 4*: With a rendering, our goal is to preserve as much unique influence of the source language as possible, while bringing it into English so that we can understand it.

> *Test development expert 3*: A rendering really is a linguistic representation of what the text would look like in the native language. A poor one, of course, but one that tries to maintain some structure, some lexical accuracy of the source language.

A rendering is supposed to not only convey the sense and content of the passage to the English-speaking test developers, but also to reflect the difficulty of the source language. Therefore, as *test development expert 3* emphasized, to accurately interpret renderings, test development experts have to avoid “looking too much at the language surface” only of the renderings. However, as all of the test development experts mentioned, sometimes, the language informant’s English skills were such that the rendering was indecipherable.

> *Test development expert 1*: So the language informants would provide translations of the passages. And sometimes I would send questions or requests for clarifications regarding those translations, and I would get the impression sometimes that they weren’t able to clarify it. Because of their limited English, that is the only way they could put it.

As for the language informants, they generally felt that renderings were a useful tool in the test development process, and made a point of providing as many details as possible to convey the full meaning of the passage to test development experts.

In light of some of the aforementioned issues, the test development experts stated that they sometimes reverted to other language informants to confirm the accuracy of the rendering.

> *Test development expert 4*: If a rendering really doesn’t make sense or is really hard to follow, sometimes it’s a better idea just to send it to someone else and have them do the rendering rather than just keep going back to the person who did it originally.

Item writing. The issues raised on passage collection also affected the next stage of test development, namely item writing (48 comments). A lot of time was spent on collecting the “right” passage so that the item writing process could go as smoothly as possible and so that the main item writer, i.e. the test development expert, would not spend time drafting a question on a part of the passage that proved to be unclear. The item writing process was described by the test development experts to typically consist of the following steps: 1) the test development expert examines the rendering and requests clarifications from the language informant, if necessary, 2) the test development expert decides on an assessment point, 3) the test development expert writes the question in English, and 4) the test development expert and language informant both review the item. The test development experts thereby stressed the critical role of the language informant, for example, in ensuring renderings of good quality to start with, in order to be able to write valid items.

> *Test development expert 2*: If a rendering is too rough and as an item writer I think, “What does this mean? How do...?” and then you try to develop a question based on that section that you don’t understand, it just doesn’t work. It’s turning into guess work. And it would become unfair to the test-taker and the integrity of the entire test if you based your questions on something that you didn’t understand. So what happens then is you have to send that back to the native speaker and say, “Please clarify, please clarify.”

The test development experts also explained that as NNSs with no (deep) knowledge of the target language and culture, it was difficult to judge whether the test content they developed corresponded to that in the target language use context (a prerequisite for test authenticity (Bachman & Palmer, 2010)).

> *Test development expert 2*: The most important step then, for that item writing process, is to have a native speaker look at your question in English against the actual language that we’re not speaking or understanding, and make sure that everything fits together.

> *Test development expert 1*: It really forces you to think that you have to do research and you have to know if you’re writing these distractors: do they, are they plausible? Do they make any sense?

Thus, liaising with the language informant was said to be indispensable at the item writing stage, for clarification and authentication purposes.

> *Test development expert 1*: And so, we’ve got that language informant to use as a resource through every phase of the process, so it can be to sort of confirm a valid assessment point, whether it’s at level or not, confirm our understanding of the passage, or a lot of times we just go ahead and draft items and send them to the language informant for review, and they will review

> the finished items and determine whether the keys are accurate and whether the distractors are plausible, but incorrect — different things like that.

This process, again, was not without its challenges, as the Hindi language informant indicated. She explained that she had to make sure to convey to the test development experts the actual difficulty of certain terms in the passage that may be the foci of certain items.

> *Language informant 2*: The item writing... It was a little bit of a problem because at times what happens is in English you just have a simple word, but that same simple word would not be a very simple word in the other language.

In sum, the participants raised a range of issues regarding the test development process for LCTLs when the assessment specialist does not speak the language being assessed. These issues will now be discussed.

Discussion

The main purpose of this study was to cast light on item writing contexts in which the assessment experts lack proficiency in the language being assessed and collaborate with language specialists. While Clark and Sundstrom-Hebert (2009) provide a rough description of such situations, the present study throws much more detailed light on the intricacies and issues of this specific collaborative item development context. Indeed, by means of interviews with both types of players, rich descriptions of issues surrounding test developers (i.e., assessment and language specialists) and the test development process were generated. This uncovered issues that are idiosyncratic to this specific test development context, but also that occur in more conventional test contexts (where the test developer is highly proficient in the target language), albeit in a less pronounced manner.

Test Developers

Item writers' expertise, selection, and training have been called "key piece[s] of qualitative validity evidence for a test" (Downing & Haladyna, 1997, p. 66). Echoing Brooks and Mackey (2009), and Hoffman et al.'s (2007) findings on LCTL testing, our research indicates that the recruitment of people proficient in a LCTL is associated with particular challenges, which means that in some cases, the profile requirements need to be loosened for this particular aspect of the test development cycle. For example, some language communities have a very small presence in the country in which the test is developed and thus only a small pool of (qualified) applicants to choose from. Or in some cases, languages are not offered in the educational system (Brecht & Walton, 1994), and its speakers, therefore, lack teaching experience (as such, or with an audience similar to the test-takers) and have limited insight into the language learning process at different levels of proficiency.

With regards to the qualifications of language testing personnel, international guidelines of good practice (e.g., EALTA, 2006; ILTA, 2007) and item writer recruitment ads (e.g. SQA, http://www.sqa.org.uk/sqa/58396.html) emphasize a number of crucial characteristics of a test writer – the key ones being 'target language proficiency and teaching experience' and 'theoretical and practical assessment expertise'. A typical test development team will thereby consist of several individuals who each possess these skills. In the LCTL case study described here, however, these desirable characteristics were distributed across the individuals who constitute the item development team: test development experts with high levels of language assessment literacy but very limited/no knowledge of the target language, and language informants who are proficient in the target language but with limited/no language assessment literacy. In order to fulfil the necessary knowledge and skill profile of an item writer, the interview data indicated that the two ideal applicant profiles in the current approach thus are: 1) a test development expert with theoretical and practical language assessment expertise, as well as foreign language teaching experience, and 2) a language informant with high proficiency in and teaching experience of the target language.

A number of additional traits were put forward by the interviewees, which are not mentioned or emphasized as much in more conventional approaches. For example, cross-linguistic and cross-cultural competencies were singled out. As is often the case in the language test development profession (Hamp-Lyons, 2000), most test development experts interviewed had previously worked as teachers of English as a second language, learned a foreign language, and lived abroad. Coupled with basic knowledge of the salient features of the target language and culture garnered through research, this instilled test development experts with global cultural competency, which is conventionally considered desirable in order to have insights into the target language and also develop tests that are suitable and appropriate for the test-takers (Kim et al., 2010). In this LCTL context, however, such competence also seems critical to the quality of interaction and collaboration with the language informants, and, by extension, most likely also the quality of the resulting test.

Similarly, language informants' international and language learning/teaching experiences were considered key to the collaboration. Ideally, the language informant is also highly proficient in the language of wider communication of the test development setting (American English in this study) to guarantee efficient cooperation. At the same time, it was pointed out that language informants should demonstrate recent and prolonged contact with the target language and culture, so that they keep abreast of changes in the LCTL, as well as socio-cultural events, in order to avoid issues related to language attrition, language interference or dated usage (cf. Brooks & Mackey, 2009; Hoffman et al., 2007). In fact, to circumvent such problems, sociolinguistic sensitivity and awareness were named as desirable qualities of both test development experts and language informants. For the language informants in particular, a valued attribute is their metalinguistic awareness, which provides them with "the ability to think about (and manipulate) language" (Gass & Selinker, 2008, p. 29). This quality is all the more crucial as language informants reflect upon their language and its singularities compared to English, and share these insights with test development experts during the test development process.

In practice, the test development experts reported some flexibility in ideal characteristics for language informants depending on the assignment. Indeed, in ‘traditional’ item writing, allocations according to strengths (and preferences) have also been reported (Kim et al., 2010). However, what was strongly emphasized in this case study – seemingly as a response to the issues raised about the qualifications of those involved in the test development – was the introduction of multiple checks in the process (see discussion below), to ensure a high-quality product.

Test Development Process

Given that language informants often lack language testing expertise (also observed by Brooks & Mackey, 2009), the first step in the test development process was basic training in proficiency assessment. The training specifically targeted practical aspects of those test development activities the language informants would work on, providing exemplars and avoiding jargon. Interestingly, item writers in Kim et al.’s (2010) study also thought less jargon was beneficial to the training even if they operated in a ‘conventional’ context and were experienced teachers. From a practical point-of-view, this targeted and restricted training seems economical. However, in the long term, it might be considered a missed opportunity of broader assessment literacy training. There seems to be particular potential in this context to extend assessment expertise and expand the international language assessment community (which is still largely dominated by experts based in particular regions and the development of tests for a limited number of target languages) to LCTLs.

After the training, the reading task development process began with passage collection by the language informants. This phase bore many similarities with the processes described in ‘ordinary’ development contexts. For example, just like in Green and Hawkey (2012), test developers would consult guidelines and topic recommendations provided by the commissioning authority before searching the Internet for materials. Also, participants in this study echoed the plea for authenticity of texts mentioned in Kim et al. (2010), which sometimes meant including “more authentic, performance-type tasks based on materials that were not specifically designed for pedagogical or testing purposes" (Leung & Lewkowicz, 2006, p. 216). When, as a last resort, language informants had to revert to purpose writing, the process seemed to match the system described by Kim et al. (2010): item writers consulting "different materials, such as newspapers, magazines, and teaching materials, to discover authentic topics and formats for reading passages" (p. 171) before creating their own reading passages for the test.

On the other hand, participants described challenges that seemed to be idiosyncratic to, or at least much more pronounced in the LCTL situation in which assessment specialists and language informants collaborate. For instance, most participants pointed out that the political situation of the target language country may severely affect the availability of passages at certain levels and the potential for breadth of topics. Also, consultation between language informant and test development expert was necessary to satisfy what Green (2014) calls "a key consideration for the developer", namely the identification of texts that both "represent an appropriate level of difficulty or challenge for the assessees" (p.105) and are of a type that "will be familiar and equally accessible to every assessee" (p.112). Renderings thereby played an important facilitating role for the collaboration between the language informant and test development expert. In addition, the matter of language interference (e.g., cognates) was brought up by all of the informants, who warned that the passage might have to be edited, or even discarded if the presence of language (variety) switching was too sizable. Finally, given the split in expertise and duties between language informants and test development experts, the language informant may have fewer insights into the actual item writing process and thus have trouble identifying texts that "will lend itself to reaching the specified number of assessment points" (Green, 2014, p. 112). Because of the aforementioned issues, extra checks were built in and language informants from different regions would cross-check the passages and renderings. In cases when the language informant was asked to semi-script a text for the purpose of the test, the main challenges were keeping the text as culturally plausible as possible and ensuring that the passage mirrored language use for communicative purposes by NSs for NSs.

The second test development phase concerned writing the comprehension questions. The general process more or less matched the one described by Spaan (2007) for ‘traditional’ settings: item writers would examine the text and write corresponding items in English before editing it, the difference being the ongoing input from language informants at each stage of the process. In that sense, the item writing process reflects that of ‘conventional’ settings in that it is “a consensus-building process within a team and an individual creative process” (Kim et al., 2010, p. 161). However, based on the test development experts’ description of the item writing process, the teamwork with language informants seems much more extensive than the individual creative process. Apart from the critical collaboration due to the split of assessment and language knowledge, an apparently higher sensitivity to sociolinguistic issues with reference to LCTLs seems to foster extra control mechanisms. For example, the test development experts seem to very thoroughly consider issues associated with language variation, standardization, and cultural appropriateness, and consult one or more language informants on these. The language informants, on their part, at times operate as reviewers of the items produced by the test development experts, thereby embedding, at least partly, a typically separate phase of the test cycle into the item writing phase (Green, 2014).

Renderings are at the center of this system of “checks and balances” and serve as the major tool for collaboration between the test development expert and the language informant. The quality of a rendering has thus strong implications for the development of test tasks. Although writing items on the basis of a rendering is not the same as translating whole tests into other languages, the quality and fairness risks associated with working with a form of translation (as, for example, discussed and shown in Ercikan, 1999; Hambleton, 2002) are likely to apply to this LCTL test development context too.

In essence, teamwork characterizes each of the three ‘typical’ phases in the work of an item writer, as identified by Salisbury (2005), i.e., the Exploratory Phase in which texts and contexts are identified, the Concerted Phase in which a first draft of the texts and items are readied, and the Refining Phase in which some form of review is followed by further polishing or revisions. Furthermore, in the present context, both test development experts and language informants alternatively took on the role of the editor at various test development stages. For example, during the passage collection stage, test development experts had the license to request edits on passages provided by the language informants; whereas during the item writing stage, the language informants could question an item written by a test development expert based on their linguistic and cultural insight. In this sense, although the indispensable collaboration between test development experts and language informants may be challenging in several respects, it seems to respond to calls for a test development process in which item writers and editors by definition work together at all stages (Baranowski, 2006). This ongoing system of checks echoes "the rigorous procedure for verification of authenticity at each step by practitioners in the field" (p. 198) that Wu and Stansfield (2001) describe in the context of LSP testing.

Conclusion

This paper concentrated on the under-researched area of item writing, which Frey et al. (2005) have argued to be ruled by collective wisdom, much more so than being underpinned by research. Through a case study approach, our research offers some insights into an even more unexplored area of language testing in which essential item writer traits are spread across different people. This item writing practice offers a componential view on language assessment literacy and challenges the traditional monolithic view of "the language test developer" by describing an instance where expertise in the target language, while indispensable, is separated from expertise in test development. The metaphor of "checks and balances", however, pertinently illustrates that such a process necessitates a continual back-and-forth between assessment experts and language informants, which serves as quality assurance at this level of development (in addition to conventional mechanisms such as use of specifications, and external item reviews at a later stage).

In spite of the richness of the data, this study was confined to parameters that risk accentuating the idiosyncrasy of its findings. Participants in other testing contexts might offer additional insights into the process of test development when test development experts collaborate with language informants. Specific issues that represent possible avenues for future in-depth research include: the influence of the language informant's English proficiency, the impact of the quality of renderings, and the role of language informants' judgment on issues related to authenticity and language (variety) switches.

With this research, it is hoped that researchers and test developers gain an understanding of the process of test development through collaboration between assessment experts and language informants. The literature on item writers/writing, especially for tests of LCTLs, is sparse, and a great deal of the theoretical and empirical work to identify and remedy constitutional and conceptual issues remains to be done. This article hopes to have laid the groundwork for such research.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Baranowski, R. A. (2006). Item editing and editorial review. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 349-357). Mahwah, NJ: Lawrence Erlbaum.

Brecht, R. D., & Walton, R. A. (1994). National strategic planning in the less commonly taught languages. *Annals of the American Academy of Political and Social Science, 532*, 190–212.

Brooks, R. L., & Mackey, B. (2009). When is a bad test better than no test at all? In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 11–23). Cambridge: 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). Bristol, UK: Multilingual Matters.

Clark, M., & Sundstrom-Hebert (2009). *Practical and theoretical considerations in proficiency test development*. Paper presented at the Confederation in Oregon for Language Teaching Fall Conference. https://casls.uoregon.edu/pdfs/conferencehandouts/assessment/TestDvptCOFLT2009.pdf

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Davies, A. (1997). The education (and training) of language testers. *Melbourne Papers in Language Testing*, *6*(1), 79–85.

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

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61–82.

Ercikan, K. (1999). Translation effects in international assessments. *International Journal of Educational Research*, 29(6), 543-553.

European Association for Language Testing and Assessment (EALTA). (2006). *EALTA guidelines for good practice in language testing and assessment*. Retrieved from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teacher and Teacher Education, 21*(4), 357-364.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). New York: Routledge.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Oxon, UK: Routledge.

Green, A., & Hawkey, R. (2012). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, *29*(1), 109–129.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*(2), 186–203.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In C. A. Porter and A. Gamoran (Eds). *Methodological advances in cross-national surveys of educational achievement* (pp.58-79). Washington, DC: National Academy Press.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, *28*(4), 579–591.

Hoffman, C., Mackey, B., Brooks, R. L., Hoffman, M., Hardy, A., Legowik, A., … Thomas, B. (2007, November). *Challenges and solutions for testing less commonly taught languages.* Paper presented at the East Coast Organization of Language Testers conference, Washington, DC.

Inbar-Lourie, O. (2012). Language assessment literacy. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 2923–2931). Hoboken, NJ: Wiley-Blackwell.

International Language Testing Association (ILTA). (2007). *Guidelines for practice*. Retrieved from http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf

Kennedy, L., & Stansfield, C. W. (2010). The Reading Proficiency Interview (RPI): A rapid response test development model for assessing reading proficiency on the ILR scale. *Applied Language Learning*, *20*, 1–16.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics. *Language Assessment Quarterly*, *7*(2), 160–174.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, *40*(1), 211–234.

NCOLCTL. (n.d.). National Council of Less Commonly Taught Languages - FAQs. Retrieved from http://www.ncolctl.org/

Richards, H., Conway, C., Roskvist, A., & Harvey, S. (2012). Foreign language teachers' language proficiency and their language teaching practice. *The Language Learning Journal*, *41*(2), 231–246.

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

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly*, *4*(3), 279–293.

Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing*, *18*(2), 187–206.

Yin, R. K. (2012). *Applications of case study research* (3rd ed.). Thousand Oaks, CA: Sage.

Zhang, Y., & Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: competing or complementary constructs? *Language Testing, 28*(1), 31-50.

Table 1 - Participants' characteristics

Participant	Gender	Native language	Experience living abroad	Experience teaching a foreign language/culture
Test development experts				
Test development expert 1	Male	English	Yes	Yes
Test development expert 2	Female	English	Yes	Yes
Test development expert 3	Male	German	Yes	Yes
Test development expert 4	Female	English	Yes	No
Language informants				
Language informant 1	Male	Farsi	Unknown	Yes
Language informant 2	Female	Hindi/Telugu	Unknown	Yes
Language informant 3	Male	Somali	Unknown	Yes

Table 2: Interview themes and code frequencies

Theme	Code	Subcode	Frequency
Test developers	Test development expert	Profile	25
		Duties	113
		Total	*138*
	Language informant	Profile	73
		Duties	126
		Total	*199*
	Relationship between the test development expert and the language informant	—	46
Proficiency scale	ILR Scale	—	77
Test development	Input Materials	Locating Material	113
		Sociolinguistics	85
		Writing System	16
		Renderings	53
		Total	*267*
	Items	Item Writing	48
Total			*775*