e-TEALS no. 5 | 2014

An e-journal of Teacher Education and Applied Language Studies

Consider the Candidate: Using Test-taker Feedback to Enhance Quality and Validity in Language Testing

David Ewing RYAN
| University of Veracruz

Abstract | This article discusses the importance for language test developers of taking candidate feedback into consideration in order to improve the overall quality and validity of their language tests. It reports on a study that was carried out at the Universidad Veracruzana, Mexico, between July and December 2010. The main objective of the study was to try and ascertain the positive and negative consequences for 245 candidates as a result of preparing for and taking a language test. The test was given in May 2010 during the spring application of EXAVER, a tiered-suite of English language certification tests developed and administered by the Universidad Veracruzana. A mixed methods approach was used for the study, incorporating both quantitative and qualitative data. The quantitative data came from the responses of a web-based questionnaire survey. The qualitative data came from the author's Research Journal, spanning a period of eight months, and from a series of semi-structured interviews. The findings of the study suggest that language test candidates not only have strong opinions (both positive and negative) about the tests they take, but they also have a strong desire to share those opinions with test developers. This type of feedback can then be used to substantially improve future tests, thereby helping to enhance the validity of the test system. The research provides a new perspective on a relatively unexplored area of language testing and has implications for language testing practitioners who welcome a more transparent and democratic form of assessment.

Key words | candidate feedback, candidate questionnaires, language test consequences, language test impact, consequential validity, collaborative language assessment, democratic language assessment, critical language assessment

Citation: David Ewing Ryan, "Consider the Candidate: Using Test-taker Feedback to Enhance Quality and Validity in Language Testing." *e-TEALS: An e-journal of Teacher Education and Applied Language Studies* 5 (2014): 1-23. ISSN 1647-712X

Introduction

In his 2004 article in *Language Assessment Quarterly*, advocating the need to "broaden, deepen and consolidate" many of our ideas about language testing, Cumming makes the convincing argument that more research is needed on the role of stakeholders in language testing contexts that have traditionally been overlooked (3). It can be successfully argued that one of these neglected areas is Mexico, and, indeed, Latin America in general. Mexico seems to be in the paradoxical situation of many Latin American countries that, on the one hand, has seen a pronounced increase over the past several decades in demand for English language instruction (and the assessment that accompanies it) while, on the other, has seen a scarcity of studies investigating the specific variables that help define the uniqueness of Mexico's context. Without doubt, one of these variables would be the candidates who actually take language tests in Mexico.

The story of these candidates, not just in Mexico but in many counties and geographic areas throughout the world, is, to a large extent, an untold one. Indeed, in the language testing literature of the last ten to fifteen years, it would be difficult to find an issue that more scholars seem to agree on than the idea that candidates are one of the most important stakeholders in language testing and yet, paradoxically, one of the most neglected ones.

Hamp-Lyons, for example, notes that "many more studies are needed of students' views and their accounts of the effects on their lives of test preparation, test-taking and the scores they have received on tests" (299). Shohamy perceives that "it is through the voices of test takers who report on the testing experiences and consequences that the features of the use of tests can be identified. Yet, in the testing literature, test takers are often kept silent; their personal experiences are not heard or shared" (*The Power of Tests* 7). And Cumming maintains that "serious consideration of the uses of language assessment requires adopting research methods that investigate people's attitudes, beliefs, cultural values, and ways of interacting . . . . Such inquiry is indispensable for understanding why people perform the ways they do in language assessment, and thus necessary for validation" (9).

The purpose of this study, therefore, was to give free rein to what can be considered as the neglected voices of test candidates in one specific test project, in one specific place, and at one specific time. The principle reason for doing this is grounded in the concept of professional responsibility that comes with being a language test developer. As McNamara and Roever insist, "language testing has a real impact on real people's lives" (8). This impact starts with the stakeholders who are immediately affected by the test, such as the test candidates and test developers, and extends outward to society at large. This impact, in turn, implies a significant amount of responsibility on the part of test developers to ensure that the tests they write and administer are as valid and reliable as possible.

One of the most valuable techniques for helping test developers to measure the validity of their tests is, precisely, by listening to the voices of candidates. Candidate perceptions, feelings, points of view, attitudes, opinions and suggestions, taken together, can serve as evidence of the positive and negative consequences of tests. In addition, feedback from candidates can serve as the impetus for discussions that can, and should, be happening not just among several stakeholders, but among many (Madaus, qtd. in Shohamy, *The Power of Tests* 149). Enlarging the dialogue in this way can help further promote not just the validity of individual tests, but the validity of the language test system as a whole, which, according to Shohamy, needs to continually "encourage testers, teachers, test takers, and the public at large to question the uses of tests, the materials they are based on and to critique the values and beliefs inherent in them" (*The Power of Tests* 131).

The article is divided into seven parts. Part 1 contains a literature review and theoretical overview of consequential validity, or the concept of looking at language tests in terms of the positive and negative consequences that are sustained by the candidates who take such tests. Part 2 gives a brief summary of the goal of the study. Part 3 contains a summary of the EXAVER English language certification tests, which served as the practical context of the study. Part 4 explains the methodology used in the study. Parts 5 and 6 offer, respectively, an overview of the findings and a discussion of those findings. Finally, Part 7 offers some general conclusions about the topic at hand.

1. Theoretical Context of Study: Consequential Validity or the Social Side of Language Testing

1.1. Scholarly Interpretations of Consequential Validity

Traditionally, consequential validity in language testing has been seen as a type of validity theory that attempts to measure the consequences of the way in which the scores from language tests are interpreted and used (Davies et al. 131). These consequences are also known as effects or impact. One of the first scholars to discuss the concept was Spolsky who reminded language testers about the important consequences their decisions may have on the lives of test candidates. He therefore urged testers to ensure that the evidence they present about the inferences they make regarding candidates' test scores is as credible and compelling as possible (Spolsky, qtd. in Bachman, "Building and Supporting a Case for Test Use" 5).

Cronbach was one of the first scholars to enlarge the concept of consequential validity to embrace society at large. He felt that the judgments that language testers make about the positive or negative consequences of a test were highly influenced by society's views of what is desirable or undesirable in the specific cultural context in which the assessment takes place. He also felt that these societal views are not stagnant but change over time (Cronbach, qtd. in McNamara and Roever 11). Cronbach's belief in the importance of social and cultural values in language testing is an essential one and was taken up by other theorists such as Messick, and Bachman and Palmer, as will be seen subsequently.

Along with Spolsky and Cronbach, another scholar who contributed immensely to the current understanding of test consequences was Messick, who defined consequential validity not as a theory in and of itself, but rather as one of the six defining aspects of construct validity (Messick, "Validity"; "The Interplay of Evidence and Consequences in the Validation of Performance Assessments"; and "Validity and Washback in Language Testing"). Messick's motivation for describing consequential validity in this way might be best explained by his definition of test constructs as being the embodiment of social values (McNamara 334).

Following Messick's important contributions to the understanding of consequential validity, the next major treatment of the topic was by Bachman, in *Fundamental Considerations in*

Language Testing. This contains a detailed section devoted to "the consequential or ethical basis of validity" (279) in which the author echoes Cronbach's and Messick's thoughts above on the important role that values play in test impact and that these values (or considerations) "are essentially political, that they change over time and they will differ from one society to another" (280).

Bachman and Palmer include an entire section in their book exploring the impact that language tests have on test candidates in particular. The authors identify three ways that language tests have a direct impact on candidates: 1) as a result of candidates' preparing for and taking the test;[1] 2) from the type of feedback that candidates receive about their performance on the test; and 3) from the decisions that testers make based on candidate scores (31).

More recent scholars, such as McNamara and Roever, and Shohamy, have also written extensively on the topic of consequential validity. McNamara and Roever wrote an entire volume that is devoted to exploring the various social considerations involved in language testing and that develops the authors' belief that a language test with positive psychometrical qualities does not necessarily mean that it will have positive social consequences (2). Shohamy is undoubtedly one of the most passionate supporters of fairness and ethics in language testing; regarding consequential validity, she stresses the need for researchers to carefully describe the nature of test consequences. She also stresses, however, the challenges and complexities that researchers may encounter in doing so, since these consequences are often invisible due to the fact that they tend to take place "outside the domains which the researcher examines" (*The Power of Tests* 49).

Shohamy also joins a fairly long list of other scholars (see Alderson and Wall; Bailey; Cheng, Watanabe and Curtis; Hamp-Lyons; Hughes; Messick, *Validity and Washback*; Wall, *Impact and Washback*; and Wall, *The Impact of High-Stakes Examinations* to name just a few) who argue that another important area of test consequences is washback, which can be defined as "the effect of testing on teaching and learning" (Hughes, qtd. in Bachman and Palmer 30). In language testing, a good example of positive washback is that identified above by Bachman and Palmer when they discuss the feedback that testers and candidates should, ideally, receive from

each other about the test, and this idea of reciprocal or mutual feedback shows up repeatedly throughout Shohamy's work, notably in her idea of "democratic" or "collaborative" assessment.

Finally, O'Sullivan echoes the concerns of earlier scholars (notably those of Cronbach, Messick, Bachman, and McNamara and Roever) who focused on the social values that are implicit in the constructs that inform language tests. According to O'Sullivan, from the very beginning of the test development process onward, test developers need to play close attention to the particular characteristics (whether they be individual, linguistic, cultural, social, or a combination of these) of the population that will be using the test, and in so doing test developers "are actually taking into consideration aspects of test consequence" (6).

For O'Sullivan, as shall be seen in Section 2.3., language tests that exhibit a high level of concern for both the local test context, and for the particular needs or realities of the candidates within that context, can be considered examples of the phenomenon he defines as language test "localization" (6).

1.2. The Business Approach

During the course of the study, the researcher decided to adopt his own approach to observing the relationship between language tests and the candidates who take them, and this was done through the lens of business administration.

It can be argued that the schools, universities, and other institutions that employ language teachers and testers, are businesses[2] in the sense that they offer:

1. a product (e.g. knowledge of a language or certification of that knowledge)
2. people who might be seen as selling the product (e.g. the teachers who work at a language school, or the test developers who write a language test)
3. people who purchase the product (e.g. the students at a language school, or the candidates who take a language test)

It also seems fair to assume that in order for any business to be successful, it needs to be aware of two key variables: the quality of the product it is trying to sell, and knowledge about the client

(or customer) base that the product is designed for. In other words, a successful business usually needs to have a well-designed and properly functioning product, but, just as important, it also needs to be familiar with and have an understanding of the requirements of the people who will eventually purchase the product.

In keeping, therefore, with the way that scholars such as Bachman and Palmer have thought about consequential validity in terms of the impact, or consequences, that language tests have on candidates, consequential validity also has to do with testers taking the time to get to know their candidates, and, more importantly, taking the time to familiarize themselves with what candidates think of the product that testers are selling them, namely language tests.

One of the best ways that testers have of ascertaining this information is through candidate feedback questionnaires that can be distributed immediately following the test administration, or, in some instances, prior to the test administration, or in still further instances, both prior to *and* following the administration, which would yield a more complete picture of candidate attitudes about the test. This information can then be used as a valuable set of qualitative data that can help compliment the quantitative data that testers receive from such elements as item analysis and descriptive statistics. While statistics are indispensable for helping to inform testers about the overall psychometric quality of their tests, candidate feed-back questionnaires can provide language testers with another way of measuring overall quality, as seen through the positive and negative consequences that candidates experience as a result of preparing for and taking a language test.

The reason of course that test developers should desire this information is so that they can identify: a) aspects of the test and the test system that seem to be working well for candidates, b) aspects of the test and the test system that seem to *not* be working well for candidates, and c) suggestions that candidates might have for improving the test and the test system. Information gleaned from any or all of these areas can then be used to substantially improve the overall quality of the test and the test system.

The topic of candidate feedback questionnaires will be discussed in greater depth in Parts 4, 5, and 6.

1.3. Importance of the Language Test System

Loosely defined, a language test system can be seen as the collective entity of all the elements that are exterior to the test *per se*, but are still indirectly related to the test. These elements might also be referred to as the "peripheral aspects" of the test, and they include, among other things, such variables as:

- **the test registration process** or how candidates are able to enrol for the test. In the case of the EXAVER tests to be discussed in Part 2, as well as for many other language certification tests, the registration process is completed online via a website.
- **the test orientation and preparation process**, which includes all the information *about* the test, including practice tests that candidates can, and should, take in preparation for the test. Again, in the case of EXAVER, all or most of this process is online.
- **the test "reception" process**, or the way that candidates are physically greeted and treated both prior to and during the test by the examiners (or invigilators) who administer the test. It is important for testers to know, for example, whether the behaviour of the examiners was calm, welcoming, and impartial, or nervous, rude, and biased.
- An example illustrating why these different elements of the test system are important can be seen in the work of Bachman and Palmer. They suggest that one way of promoting positive test impact is by involving candidates at various stages of the test development process and by soliciting candidates' feedback on such things as their overall impression of the test, as well as their impression of more detailed aspects such as the tasks that appear in the test. Bachman and Palmer argue that if candidates are involved in this way, it is very possible that they will have a more positive perception of the test, will have more confidence and motivation when taking it, and, as a result, will very likely perform better (32).

The same argument might be made for why candidates should be highly encouraged to take a practice test, which, as noted above, can be considered to be part of the "test orientation and preparation process". If candidates perceive the tasks in the practice test as being fair and of good quality (or, in the words of Bachman and Palmer, as being "authentic" and "interactive"), then, according to Bachman and Palmer's hypothesis, this would likely serve to increase candidates' sense of confidence and motivation regarding the practice test, which could, likewise, help them to perform better on a live test. By the same token, but related to the "test reception process", if candidates are treated with respect and courtesy by invigilators on the day of the test, this could help them to feel more at ease, and to concentrate better, which, once again, might help them to perform better. The implications for test validity are clear: testers obviously want candidates to perform to the best of their ability, so that the scores calculated based on candidate performance are as fair, valid, and authentic as possible.

To conclude, therefore, elements of the test system can be considered to be related to the validity of the test, in the sense that if these elements are sufficiently attended to, this could help candidates to form a positive impression of the test and to perform to the best of their ability, which would, in turn, enhance the overall validity of the test. But if these elements are not attended to, then the opposite scenario might occur and candidates could form a negative impression of the test which could, then, interfere in their performing to the best of their ability, which would likewise diminish the overall validity of the test.

2. Goal of the Study

The study focused specifically on what Bachman and Palmer considered to be one of the three ways that language tests have a direct impact on test candidates, namely, in terms of the consequences that candidates experience as a result of preparing for and taking these tests (31). In order to measure this impact, it was necessary to liberate the voices of the test candidates who participated in the study, and this, then, became the primary goal of the study. This was accomplished, first, by soliciting candidates' opinions both about the test they took and about the

test system, and second, by soliciting their suggestions of ways to improve the test and the test system.

3. Practical Context of Study: The EXAVER English Language Certification Tests

3.1. General Description

EXAVER is the name of the tests that were used as the basis of the study, and refers to a tiered-suite of English language certification tests that are administered by the Universidad Veracruzana (UV), in the south-eastern Mexican state of Veracruz. The suite was developed in the year 2000 by a small group of English language teachers from the UV, as well as representatives from the following international organizations: The British Council, Cambridge Assessment, and Roehampton University's Centre for Language Assessment and Research (CLARe).

The construct behind the EXAVER tests is to measure three language proficiency levels that are informally linked to the Council of Europe's *Common European Framework of Reference for Languages* (CEFR) in that each EXAVER level is based on the Council of Europe content specification that inspired the corresponding CEFR level (Council of Europe 23). This is summarized in Table 1 below. The EXAVER tests are administered twice a year, once in the spring and once in the fall, at 11 separate language centres throughout the state of Veracruz.

EXAVER	CEFR	Council of Europe
1 **Upper Beginner**	**A2**	**Waystage**
2 **Lower Intermediate**	**B1**	**Threshold**
3 **Upper Intermediate**	**B2**	**Vantage**

Table 1: *Levels of EXAVER tests and their corresponding CEFR L* (adapted from Abad et al.)

3.2. Test Structure

Each EXAVER test contains three separate sections, or papers, and the structure of each of these sections is described in Table 2.

Paper 1 Reading and Writing	Paper 2 Listening	Paper 3 Speaking
• 5 parts • Variety of tasks: matching, multiple choice, modified cloze text • Indirect measure of writing	• 4 parts • Range from comprehension of relatively short informal conversations to comprehension of more formal and substantially longer conversations	• 3 parts • Combine some type of interview task (interlocutor to candidate), discussion task (between a pair of candidates) and a long-turn task (interlocutor to candidate)

Table 2: *EXAVER test structure* (after Dunne)

3.3. Test Localization

According to O'Sullivan, one of the defining characteristics of the EXAVER examinations is that they represent "the first systematic attempt to create a 'local', affordable, and sustainable language test system" (10). In focusing their attention on the local geographic context where the examinations take place (e.g. south-eastern Mexico), and on the particular needs of the test candidates within that context (students, primarily, of the Universidad Veracruzana), EXAVER's test developers helped create a process now known as "localization". O'Sullivan defines this as "the practice of taking into account those learning-focused factors that can impact on linguistic performance . . . [and] the recognition of the importance of test context on test development . . ." (6).

Economic affordability was one of the first local variables that EXAVER's test developers took into consideration. Due to the fact that the majority of EXAVER's candidates were (and continue to be) unable to afford the cost of more reputable international English language certification tests, EXAVER's test developers decided to create a suite of economically affordable tests, more in line with median to lower income brackets based on the Mexican minimum wage.[3]

Table 3 shows the current costs (as of September 2014) of taking an EXAVER test, with their approximate equivalents in Euros.[4]

LEVEL	Cost in MX Pesos	Cost in Euros
EXAVER 1	350	Approx 21
EXAVER 2	400	Approx 23
EXAVER 3	450	Approx 25

Table 3: *Comparative cost of taking an EXAVER test*

4. Methodology

4.1. Type of Data and Participants

The study included both quantitative and qualitative data. The quantitative data came from the responses of 245 EXAVER candidates who completed a web-based questionnaire survey, administered in the summer of 2010, following the spring 2010 test administration of EXAVER's three levels. The qualitative data came from the author's Research Journal, spanning a period of eight months, from March to October 2010, and from a series of semi-structured interviews conducted in October 2010 with four of the questionnaire's respondents.

4.2. Type of Method

A mixed methods approach for data collection and analysis was used for the study, starting with a quantitative investigation in the form of a web-based questionnaire and then followed by a qualitative study, in the form of semi-structured interviews. In practical terms this meant that at the end of the questionnaire, participants were able to tick a box and include their name and email address, signifying their desire to be contacted by the researcher for the second phase of the study.

4.3. Data Collection & Analysis of Quantitative Phase[5]

The web-based questionnaire included 44 questions, comprising 42 closed-format, multiple-choice questions, and two open-ended questions. Of the closed-format questions, 10 employed a Likert Scale, with options spanning from 1 to 5, as a way of ascertaining candidates' opinions or feelings about a variety of topics related to the test they took and to the test system. Excel Version 2003 was used to analyse the data.

5. Findings

Out of the total 964 candidates who took an EXAVER test in May 2010, 245 of them (or 25%) responded to the survey. Of these 245 candidates, 99 (40%) ticked the box at the end of the survey demonstrating their desire to participate in the semi-structured interviews that constituted the second phase of the study. This, therefore, was the first finding of significance, namely, that such a large percentage of candidates wished to participate in the second phase of the study and further elaborate on their opinions about the process of preparing for and taking the test. This indicated an apparent high level of interest among EXAVER's candidates to have their voices heard.

Web-based questionnaire survey

As research instruments, questionnaires, like all forms of data collection, have their own distinct advantages and disadvantages. In terms of the latter, researchers sometimes complain about the lack of depth and richness that is found in multiple-choice responses (Dörnyei 115). For this reason, the researcher chose to include two open-ended questions in the survey, along with the overwhelming majority of 42 multiple-choice questions. While the responses to all of the survey's questions provided important feedback, the responses to the two open-ended questions (identified in the survey as numbers 17 and 30) are noteworthy, due both to the high number of candidates who responded to them (well over half of the total 245 candidates who responded to the survey), as well as to the diversity of their answers. Summaries of these responses follow.

Question 17

The text for question 17 read:

> *"Do you feel that there is anything we could include on the EXAVER website that might help future candidates to feel less anxious and/or more confident before taking the test? If so, please write your comment(s) below, taking all the space that is necessary."*

Question 17 yielded 144 total responses, including:

- 23 positive comments, such as:
  - *"The teachers who administered the test were excellent, and they worked well together as a team."*
  - *"I didn't hire a tutor or use any books to prepare for the test, but I found the information on the website very useful."*
  - *"Everything on the website is very clear – congratulations!"*
  - *"EXAVER is an excellent alternative for certifying your level of English, and it's great that it was developed here at our university."*
- 24 negative comments, such as:
  - *"The waiting time to get your grade is too long... you really need to find a way to make it go faster."*
  - *"The noise from traffic in the street at the language centre where I took the test made it very difficult to hear the CD during the listening section."*
  - *"My chances of passing the test would have been much better if the preparation materials had been more accessible; I received an 'error' message when I tried to open the Sample Tests on the website."*
  - *"I would have benefited from a greater variety, and greater scale of difficulty, of test preparation materials – the Sample Tests on the website were really easy and not very helpful."*
- 97 suggestions, such as:
  - *"Include a video on the website of a sample Speaking Test".*

- *"Include testimonies or opinions on the website from past candidates who had a positive experience taking the test."*
- *"Include a bibliography on the website of literature to consult for helping candidates to prepare for the test."*
- *"Include a description (either on the website or included with the diploma) of how grades are calculated."*

Question 30

The text for question 30 read:

> *"Do you have any other comments (positive or negative) and/or suggestions that you'd like to add regarding the EXAVER test you took or about the EXAVER Project in general? If so, please write them below, taking all the space that is necessary."*

Question 30 yielded 127 responses, including:

- 38 positive comments, such as:
  - *"I liked the test – everything seemed very clear and precise. Thanks."*
  - *"The EXAVER staff appeared to be very knowledgeable and when they gave the instructions in English, it was very clear, which set me at ease and made me feel more confident."*
  - *"I feel lucky for having had the opportunity of taking an EXAVER test and it was a great experience."*
  - *"Everything was fine, and the level seemed very appropriate."*
- 61 negative comments, such as:
  - *"It was very tedious waiting so long to take the Speaking Test."*
  - *"It was frustrating having to physically go to the language centre where I took the test in order to get my diploma – they only give them out in the mornings, and I disagree with this policy."*

- *"While waiting in line to enter the test centre, I was told that my name was not on the list even though I had my registration receipt. In the end I was able to take the test, but I felt very nervous."*
- *"The pencil they gave me for filling in the Answer Sheet was of really poor quality."*

- 28 suggestions, such as:
  - "*It would be nice to have a more detailed report on how I fared on the test, such as knowing how I performed on each part of the test maybe in terms of percentages."*
  - *"In order to accommodate the needs of students, there should be more applications of the tests than just twice a year."*
  - *"You should design a course of several months duration that students could take for helping them to prepare for the test."*
  - *"There should be more publicity for the tests, especially for those of us who are not students of the Universidad Veracruzana, but rather from the community at large."*

6. Discussion

6.1. Specific Concerns

The phrasing of Question 17 in the web-based survey, with special emphasis on the words "more confident" and "less anxious", was intentional as a way of reflecting the researcher's premise[6] that the less anxious and more confident candidates feel before taking a language test, the more likely they are to perform better. The relatively long list of suggestions (97 in total) that candidates gave in response to this question have, therefore, proven quite useful in helping EXAVER's test developers and administrators to improve the quality of the preparation materials on the exam board's website so that candidates can, indeed, feel more confident and less anxious before taking a live test.

Question 30 in the web-based survey should seem familiar to researchers, since it is the classic *"Do you have anything else to say?"* type of query that is usually included as the final

question in an oral interview. It was considered necessary to use it as an open-ended question in the survey as a type of "safety net" in order to ensure that candidates were given the opportunity of stating anything, and everything, they wished to state about the process of preparing for and taking an EXAVER test.

One of the negative responses to Question 30, referring to a candidate's sense of anxiety over their name not being found on the official list of registrants for the test, relates to the theme of Question 17. It should serve to remind testers of the importance of taking measures to avoid circumstances that might create unnecessary stress or anxiety for candidates on the day of the test. One way of doing this is for test examiners and administrators to meet together and develop a list of all the things that could feasibly go wrong on the day of the test, and then to come up with an effective way of dealing with each of them. Each potential problem and its corresponding solution could then appear on a printed sheet of paper that could be given to invigilators on the day of the test.

By contrast, one of the positive responses to Question 30 illustrates how what might be interpreted as a rather routine, mundane task (reading the initial instructions once candidates are seated) can actually serve to minimize stress and anxiety, and to boost candidates' sense of confidence, provided the instructions are read calmly and clearly. Both of these examples serve to reinforce the importance of ensuring that the test "reception" process (see Section 1.3.) is as smooth and professional as possible.

6.2. General Concerns

Candidate responses from both the questionnaire survey and the semi-structured interviews provided a rich representation of the diversity of opinions, feelings, perceptions, and attitudes that EXAVER candidates have about the tests they take. They also provided EXAVER's test developers with important insight regarding some of the positive and negative consequences for test candidates as a result of preparing for and taking a language test. With particular regard to the questionnaire survey, the quantity and variety of responses bring to mind Shohamy's

observation that the overwhelming majority of test candidates not only have a strong need and desire to express their feelings about the test they took, but they also have the inherent *right* to do so, and it is the responsibility of language teachers and testers to *enable* them to do so (*The Power of Tests* 156). By providing for this, she feels that testers can help democratize the act of taking a test so that the experience becomes more of a collaborative, horizontal process, rather than an authoritarian, top-down one (136-37).

It can be argued, however, that the most important step that takes place in the overall process of soliciting candidate feedback is what testers finally end up doing with this feedback after receiving it. For this reason, one might correctly refer to the "final consequences" of consequential validity, for it is the final actions that test developers take regarding candidate feedback that could serve to increase the likelihood of positive consequences occurring for future candidates and, accordingly, could serve to decrease the likelihood of negative consequences occurring for those candidates.

As a way of illustrating how a language test board can convert candidate feedback into concrete actions that will hopefully generate positive impact for future candidates, the following is a list of actions that EXAVER has already undertaken or is currently undertaking based on candidate feedback from this and other studies:

- Streamlined registration process, making it much easier for current and future candidates to register for the tests.
- New online grade allocation process to substantially reduce the waiting time for receiving grades.
- Sample Speaking Test for each of EXAVER's three levels, uploaded to the EXAVER website so that potential candidates have an idea of the format of the test, as well as the type of tasks they can expect to encounter. These tests serve to compliment the Sample Reading, Writing, and Listening Tests that have appeared on the website since EXAVER's inception.

- Drafting of a document with a list of administrative procedures that can potentially be problematic for examiners and invigilators on the day of the test, along with their corresponding solutions.
- Dissemination of candidate feedback questionnaires in order to continue to monitor the positive and negative consequences for the candidates who take the tests.
- Analysis and discussion of appropriate action(s) to take based on candidate responses to the questionnaires.
- Follow-through to confirm that appropriate action was in fact taken.

7. Conclusion

By now it has perhaps become apparent to the reader that what candidate feedback and consequential validity in language testing actually relate to is a type of assessment that is more inclusive and democratic in nature than the traditional, authoritarian type of model that was prevalent in so many assessment contexts throughout the world during much of the twentieth century and, indeed, prior to that.[7]

When test developers refuse to solicit candidate feedback, or do so without following through on it, this only serves to reinforce the undemocratic nature of the assessment, and the power and control that testers often exert over the candidates who take their tests. Conversely, when test developers solicit candidate feedback and then take positive actions based on that feedback, this serves to strengthen the overall democratic nature of the assessment by revealing a horizontal, collaborative process. Moreover, this process encourages the participation of not merely a few, but a wide variety of stakeholders, thereby strengthening even further the democratic nature of the process.

Another important point that language test developers should consider when judging the validity of their assessments is that language testing, like any type of testing, is, at best, an inexact science. There is an innumerable amount of things that can go wrong on the day of the test and that can interfere in its validity. The air conditioning in a hot and humid room could suddenly stop

working, thereby forcing candidates to finish the remainder of the test in uncomfortable physical conditions. Or an oral examiner could ask a candidate what s/he did on her last vacation, without knowing that someone in the candidate's immediate family died at that time. In both of these not overly extraordinary cases, the candidate's concentration could feasibly be thrown off, thereby negatively affecting his/her performance on the test, which would also mean that the score the candidate receives on the test is not a true reflection of his or her ability.

The above examples represent real situations that have taken place during previous EXAVER test administrations. Due to the fact that language testers work with real people in the real world, real problems are bound to occur, and there is very little that testers can do to ensure that these problems will no longer occur in the future. There are, however, many things that language testers are in fact able to control when it comes to designing and administering their tests. These include the following:

- Concern for the test's most important stakeholder: the candidate.
- Collective elements of the test system such as the test registration process, the test orientation and preparation process, and the test reception process.
- The overall quality of the test *per se*, e.g. its reliability, and validity of construct.
- Being responsible and effective examiners, e.g. giving fair and non-partial treatment to all candidates and following-up after the test by writing a "post-exam" report with a list of things that went right and wrong during the test application.

By concerning themselves with these and other important variables, language testers can help safeguard the overall fairness and integrity of the test and the test system. In so doing, they also help to underscore the difference between assessments that, on the one hand, are moving towards a more dynamic, responsible, and democratic model, and on the other hand, ones that continue to remain more stagnant and conventional in nature.

Notes

[1] As shall be seen in Part 2, the focus of the study was on this first type of impact described by Bachman and Palmer.

[2] The use of "business" here stems *not* from an interpretation of the word focused on such variables as volume and profitability, but rather, in a more general sense, as a synonym for a place providing an exchange of goods.

[3] As of December 2014, the Mexican minimum wage was approximately 61 pesos per day.

[4] For more details on the EXAVER examinations and the EXAVER test system, especially as they relate to localization, see Abad et al. "Developing affordable, 'local' tests: the EXAVER Project," in *Language Testing: Theories and Practices*. Ed. Barry O'Sullivan. Palgrave Macmillan, 2011. 228-43.

[5] Due to space considerations, information related to the methodology and findings from the second (qualitative) phase of the study could not be included here, but is available by contacting the author at <dewing@uv.mx>.

[6] This premise was itself based on Bachman and Palmer's similar hypothesis. See Section 1.3.

[7] The traditional or authoritarian model of education and assessment is of course still prevalent in many parts of the world today, including in many educational contexts in Mexico.

Works Cited

Abad, Adriana Florescano, B. O'Sullivan, C. Sanchez Chavez, D. Ewing Ryan, E. Zamora Lara, L. A. Santana Martinez, M. I. Gonzalez Macias, M. Maxwell Hart, P. Evelyn Grounds, P. Reidy Ryan, R. A. Dunne, and T. de Jesus Romero Barradas. "Developing Affordable 'Local' Tests: the EXAVER Project." *Language Testing: Theories and Practices*. Ed. Barry O'Sullivan. Oxford: Palgrave Macmillan, 2011. 228-43.

Alderson, Charles, and Dianne Wall. "Does Washback Exist?" *Applied Linguistics* 14.2. (1993): 115-29.

Bachman, Lyle. *Fundamental Considerations in Language Testing.* Oxford: Oxford UP, 1990.

- - - . "Building and Supporting a Case for Test Use." *Language Assessment Quarterly* 2 (2005): 1-34.

Bachman, Lyle, and Adrian Palmer. *Language Testing in Practice.* Oxford: Oxford UP, 1996.

Bailey, Kathleen. "Washback in Language Testing." *TOEFL Monograph* 15. Princeton, NJ: Educational Testing Service, 1999.

Cheng, Liying, Yoshinori Watanabe, and Andy Curtis. *Washback in Language Testing: Research Contexts and Methods.* Mahwah, N.J.: Lawrence Erlbaum & Associates, 2004.

Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching and Assessment.* Cambridge: Cambridge UP, 2001.

Cumming, Alister. "Broadening, Deepening and Consolidating." *Language Assessment Quarterly,* 1 (2004): 5-18.

Davies, Alan, A. Brown, C. Elder, K. Hill, T. Lumley, and T. McNamara (1999). "Dictionary of Language Testing." *Studies in Language Testing.* Vol. 7, Cambridge: University of Cambridge Local Examinations Syndicate, 1999. 118-19.

Dörnyei, Zoltán. *Research Methods in Applied Linguistics.* Oxford: Oxford UP, 2007.

Dunne, Roger. "The EXAVER Project: Conception and Development." *MEXTESOL Journal,* 31.7 (2007): 23-30.

Hamp-Lyons, Liz. "Washback, Impact and Validity: Ethical Concerns." *Language Testing* 14 (1997): 295-303.

Hughes, Arthur. *Testing for Language Teachers.* Cambridge: Cambridge UP, 2003.

McNamara, Tim. (2001). "Language Assessment as Social Practice: Challenges for Research." *Language Testing* 18 (2001): 333-49.

McNamara, Tim, and Carl Roever. *Language Testing: The Social Dimension*. Oxford: Blackwell Publishing, 2006.

Messick, Samuel. "Validity." *Educational Measurement*. Ed. R.L. Linn**.** 3rd ed. New York: American Council on Education, 1989. 13-103.

---. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23 (1994): 13-23.

---. "Validity and Washback in Language Testing." *Language Testing* 13 (1996): 241-56.

O'Sullivan, Barry, ed. *Language Testing: Theories and Practices.* Oxford: Palgrave Macmillan, 2011.

Shohamy, Elana. "Testing Methods, Testing Consequences: Are They Ethical? Are They Fair?" *Language Testin,* 14 (1997): 340-49.

---. "Democratic Assessment as an Alternative." *Language Testing* 18.4 (2001): 373-91.

---. *The Power of Tests: A Critical Perspective on the Uses of Language Tests.* London: Longman, 2001.

Wall, Dianne. "Impact and Washback in Language Testing." *Encyclopedia of Language and Education*. Eds. Caroline Clapham and David Corson. Vol.7: Language Testing and Assessment. Dordrecht. The Netherlands: Kluwer Academic Publishers, 1997. 291-302.

---. *The Impact of High-Stakes Examinations on Classroom Teaching, a Case Study Using Insights from Testing and Innovation Theory.* Cambridge: Cambridge UP. 2005.