Content list available at www.urmia.ac.ir/ijltr

Iranian Journal of Language Teaching Research

Urmia University

The effect of test specifications review on improving the quality of a test

Hamed Zandi [a, *], Shiva Kaivanpanah [b], Seyed Mohammad Alavi [b]

[a] *University of Tehran & Institute for Advanced Studies in Basic Sciences (IASBS), Iran*

[b] *University of Tehran, Iran*

ABSTRACT

Reviewing the test specifications to improve the quality of language tests may be a routine process in professional testing systems. However, there is a paucity of research about the effect of specifications review on improving the quality of small-scale tests. The purpose of the present study was twofold: how specifications review could help improve the validity of a test in the context of assessment for learning (AFL) and to what extent qualitative review of items can identify poor ones. To this end, a group of trained test specifications reviewers (*No = 7*) provided feedback on the specifications of a test and the quality of the items. Analysis of feedback showed that pedagogical concerns naturally emerged during the specifications review and led to concrete suggestions on how the specifications could be revised so that the resulting test could become more useful in a classroom context. Moreover, the test items were administered to a group of (*No = 40*) test-takers and a set of quantitative item analyses was carried out. Comparison of the results of qualitative judgment of reviewers with the quantitative analyses showed about 38 % of the items suggested for revision by the reviewers were also identified as poor by the classical test theory (CTT) analysis. The findings highlight the potential of specifications review as part of the a priori validation of tests in small-scale assessments where conducting statistical analysis is not usually feasible.

Keywords: test specifications; feedback from reviewers; reviewer characteristics; a priori validation; small-scale tests

ARTICLE SUMMARY

Received: 5 Feb. 2013 **Revised version received:** 29 Nov. 2013

Accepted: 1 Dec. 2013 **Available online:** 1 Jan. 2014

* Corresponding author: University of Tehran & IASBS, Iran
Email address: zandi@iasbs.ac.ir

10.30466/ijltr.2014.20420

Introduction

Designing test specifications has been considered as one of the core processes of test development (Bachman & Palmer, 1996/2010; Fulcher & Davidson, 2007; Saville, 2012). One reason for their importance is that specs have the role of a generative blue print, to use an engineering metaphor, from which many equivalent test items or tasks can be produced (Davidson & Lynch, 2002). To that end, specs undergo cycles of revision and feedback and proceed through successive versions before a satisfactory interpretation of the construct and equivalency of items is reached. Furthermore, specs are consensus-based i.e., they are developed by a team of test developers and item writers working together and accommodating each other's views (Davidson, 2012). However, there are not many studies in the literature that empirically demonstrate whether and how specs review contributes to the quality of a test. Similarly, the extent to which qualitative review of test items can be helpful in detecting problems with items is still an open question for research.

This study explores the utility of reviewing the specs of a diagnostic test of grammatical knowledge intended to be used in the context of AFL. It collects feedback from trained specs reviewers to find whether pedagogical considerations naturally arise during the review process of a test that will ultimately be used in a classroom. Moreover, it gathers empirical evidence to compare the results of qualitative judgment of spec reviewers with those of quantitative analysis of items in identifying poorly written ones. The findings can be used to support the usefulness of an a priori validity argument for small-scales tests.

The majority of reported work in language testing is heavily entrenched in quantitative analysis of test scores. However, this study was a step in the general direction of accumulating qualitative evidence obtained a priori that can complement the more routine quantitative line of validity argument that is typically constructed a posteriori.

Background to the study

Since Messick (1989) proposed both "empirical evidence and theoretical rationales support the appropriateness and adequacy of inferences" (p. 13) the field has moved to underscore the process of validation (Chapelle, 2012).

In this regard, Weir (2005) proposes that an evidence-based approach to validation can be vetted by two types of evidence: a priori evidence and a posteriori evidence. A priori evidence is collected during the design of the test and includes theory based validity evidence, (e.g. defining the construct of reading comprehension and context validity evidence, that is, task setting, task demand, setting and administration, and characteristics of test takers). A posteriori evidence is gathered after the test is administered and consists of scoring reliability and validity, criterion-related validity, and consequential validity. Thus, for Weir (2005), validation starts from the first step of test design and at those early stages it deals with collecting evidence regarding abilities the test is supposed to measure and how sample of tasks can measure those abilities.

Language testing literature is generally silent on the process of specs development and its potential for a priori validation of a test. Davidson (2012) maintains that reviewing specs of tests and improving them is in line with current validity theories in language testing that suggest validity issues should be considered from the early stages and in all stages of test development (Chappelle, 2012; Weir, 2005). Similarly, Saville (2012) proposes that in order to increase the quality of any assessment the entire process in test development needs to be subjected to two types of control procedures: quality control and quality assurance. The former refers to adherence to a series of

guidelines by the test developers themselves and the latter refers to an external review to monitor the quality. This view resonates well with the literature on specs as preparing them is a process of review and feedback. Spec documents are metaphorically regarded as organisms that evolve. This evolution from an existing version of the specs to the next version is facilitated by scrutiny and feedback from reviewers (Davidson & Lynch, 2002; Li, 2006).

Despite these calls in the literature to utilize the potential of specs reviewing in building a stronger validity argument, only a few studies have been conducted on how specs can be audited (i.e., reviewed diachronically) to observe the factors that have contributed to changes in them (Li, 2006), and how specs are perceived by item writers (Kim, Chi, Huensch, et al., 2010).One reason for paucity of research in this regard is that, as Davidson (2008) argues, the canon of language testing is limited with a tradition that largely takes for granted issues and practices in test development and is more obsessed with statistical issues of testing. Moreover, validity research in language testing literature has mainly focused on large-scale assessment and procedures such as factor analysis, G-theory, and IRT that are not readily transferable to small to medium scale tests (Davidson, 2008). The canon exists despite the fact that the majority of tests developed in educational settings are small to medium scale and are prepared by teachers.

Underlining educational function of assessment is currently at the core of assessment reforms in different parts of the world (Inbar-Louri, 2010), where teachers are expected to have an integral plan to diagnose the weaknesses and strengths of their students, provide feedback, and adapt their teaching continuously. UK Assessment Reform Group (2002) defines assessments with the end result of learning, i.e., AFL, as "the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (p. 2–3). The definition implies that feedback to the learners plays a key role in learning process. Along the same lines, as Turner (2012) stresses, an assessment that takes place in classroom is a "contextually bound and socially constructed activity involving different stakeholder in learning" (p.66) where the main participants are teachers and learners. Therefore, the insight the teachers have about their students may play a significant role in shaping the tests they design.

The question remains how these teachers, who may not have adequate training in psychometrics, or lack the time or other resources for rigorous quality, control still argue the tests they are developing enjoy acceptable levels of quality. As there is a growing concern for this type of accountability, test developers in institutional settings and language teachers may venture paths, other than psychometric ones, to ensure the quality of their tests.

This study is an attempt to offer suggestions as to how investigating the process of development of specifications for a test can bring validity considerations to the focus right from the beginning stages of a testing project, help increase the quality of test items, and help teachers build a case for accountability. More specifically the study seeks to answer the following questions.

RQ1: Does the spec review process contribute to the a priori validation of a test in an AFL context?

RQ2: What percentage of the items that are suggested for revision by a qualitative review is also suggested for revision by a quantitative review?

Method

The participants

The participants for the spec review phase of the study were seven graduate students and TESOL teachers and a TESOL graduate. Table 1 shows the profile of the participants of the spec review. All of the participants were trained in a five-week workshop on test spec development and reviewing as part of an advanced language testing course. The sampling method was convenience sampling because it reflects the real life limitations small to medium scale test developers face, as they have inadequate resources to recruit a professional team of experts. Nevertheless, it was important to control recruiting participants to include diverse voices and perspectives e.g., those of native speakers (NSs) and nonnative speakers (NNSs), novice theory-oriented and veteran practice oriented teachers, and testing oriented and teaching oriented panelists. Therefore, reviewers for the study were selected based on their knowledge of language testing and teaching experience and diverse perspectives they brought to increase the generalizability of this investigation by taking into account more diverse views. The reviewers were compensated for their time on an hourly basis.

The test-takers for the other phase of the study were 40 male and female non-English major university students studying in an EAP course in a major research university in Iran. According to the English section of the university, the EAP course attendants' ability ranged between elementary to intermediate based on Oxford Quick Placement test. The participants, who were selected by convenience sampling, were asked to take the test and attempt all of the items. They were compensated for their time by a gift.

Table 1

Profile of the Participants in the Main Review of the Specifications for a Diagnostic Test of Grammatical Knowledge

Pseudonyms of the reviewers	Sara	Harry	Tina	Kate	Cathy	Lucy	Jane	Test coordinator
Gender	Female	Male	Female	Female	Female	Female	Female	Male
Age	28	31	32	24	22	29	48	27
Last degree	PhD in progress	PhD in progress	PhD	Master's in Progress	Master's in progress	Master's in progress	Masters	PhD in progress
Major	SLA and teacher education	Applied Linguistics	Educational Psychology (specialized in language testing)	TESOL	TESOL	TESOL	TESOL	TESOL
Native language	French	Urdu, Punjabi	Chinese	English	English	English	Mandarin Chinese	Bilingual in Azeri and Farsi
Other languages	English Spanish	Saraiki, Arabic, Pashto	English	Spanish	Spanish	Spanish	English	English Arabic
Experience of teaching	6 ½ years of FSL	5 years of ESL	6 years of ESL	1.5 year of ESL	4 years of ESL	2 years of ESL	15 years of ESL	10 years of EFL
Experience of teaching grammar	Grammar of French	Grammar of English in Pakistan	Very limited experience of teaching grammar	Tutoring grammar	Grammar as integrated in task-based approach	Integrated teaching of grammar	Mostly teaching grammar	Grammar as both integrated in Task based approach and form focused instruction
Taking an advanced course on testing	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other testing courses	Advanced Seminar in testing 2012	Advanced Seminar in testing 2012	----	No	No	No	No	Advanced Seminar in testing 2012

The instruments

The test specifications. The test specs developed and reviewed in this study was part of a testing project aimed at developing a diagnostic test for a set of grammatical structures. The test was intended to diagnose the weaknesses and strengths of students who came to a three semester long English for Academic Purposes (EAP) course at an Iranian university, where English was not the medium of instruction.

The specs for the test were written, by the main test coordinator, who had the experience of teaching in that context for some years, with adherence to guidelines provided by Popham (1978) as modified by Davidson and Lynch (2002).The specs document had the following components: general description, prompt attributes, response attributes, sample item, and specification supplement. The specs document was made of two parts. The first part included an introduction to the test and general guidelines for specs. The second part was more detailed and included the specific item writing guidelines, directions, and sample items for three grammatical categories of tense and aspect, relative clauses, and conditional sentences.

Feedback booklet. A questions booklet was prepared to obtain feedback from the participants of the review. It consisted of a number of structured and open-ended questions about different sections of the specs, the items, and how they could be improved, more specifically, whether the items were in line with the construct of the test, they needed to be modified, there were adequate number of items per grammatical category. However, there were no questions about how the test might be used in classroom or its pedagogical effectiveness could be increased.

The test items. Since the specs already included the sample items, those items appeared on the test. There were 29 editing and 32 multiple-choice (MC) items testing a wide variety of grammatical structures (Appendix).

Procedure

To increase the quality of feedback from the participants of the spec review, meetings were held with a few participants at a time instead of holding a shorter meeting where all the participants could attend together. Because it was assumed that in face-to-face meetings more information could be exchanged and the synergy produced by other participants' contribution to the group discussion could spark more insights.

Thus, to maximize the amount of one-on-one exchange of ideas, the participants were divided into four groups (three sessions with two participants and one session with one participant) and the main test coordinator met with each group separately. The meetings were in the format of a half-day workshop where the participants read the specs, discussed certain issues, and wrote their reflections in the feedback booklet.

Because of the time constrains, it was not practical for the reviewers to cover all the second part of the specs in the workshop. Thus, the second part was divided into three categories of tense and aspect, relative clauses, and conditionals each to be discussed by a different group of reviewers. Still all of the reviewers had already studied the first part that constituted 40% of specs and answered more than 70% of the same questions in the feedback booklet. Moreover, two participants, a NS and a NNS of English, reviewed and commented on both parts of the specs focusing on the sample items.

After the first part of the workshop dealing with general descriptions was completed, the participants started writing in the feedback booklet. They read the questions in the feedback booklet, read the relevant part of the specs in the specs booklet, discussed with each other, and finally wrote their reflections on the questions in the feedback booklet.

All the discussions were voice recorded and transcribed for further analysis. After the feedback from the reviewers for the second part of specs was gathered, and the voice recordings were transcribed and compared to the written comments, all the answers to each category of questions were synthesized. Afterwards, the test, that was comprised of the items on the specs document, was prepared and administered to the test-takers who were told to attempt all the items. The purpose of the quantitative item analysis was to detect potential problems with items. Therefore, the test papers were scored and data was analyzed using Excel and SPSS. Item level statics such as item facility, item difficulty, point-biserial correlation, Cronbach's alpha, and distractor analysis were performed.

Results and discussion

The first research question was about the contribution of spec review on the a priori validation of a test in an AFL context. In this regard, all the participants read the relevant part of the specs i.e., the beginning part of the specs document that provided some general descriptions of specs including the context of the test use, purpose of the test, and the item types used.

More specifically, the participants, in different sessions, studied the general descriptions on their own, asked some questions for clarification, and offered suggestions on how that part of the specs could be improved. The comments could be classified to eight categories: presentation of the guiding language, specifying the audience for the specs, need for more background information about the test-takers and testing situation, need for clarification of certain terms, item formats, scoring, feedback to the test-takers, and the role of construct irrelevant factors such as vocabulary knowledge in a test of grammar. Of all the eight categories, three seem to be more interesting: background information, audience, and how to give feedback relevant to the pedagogic aspect of test use. Below is a summary of the questions that occurred to the participants:

> How will the test be used in class?
>
> What are the goals of the undergraduate and graduate EFL courses in the context of that University?
>
> What is the nature of the courses for which the test will be used?
>
> What is the knowledge the students are supposed to have come with but they do not?
>
> What is the syllabus to be covered in the course?

As it can be seen most of the questions were related to test use in a teaching context. Even before the participants read a prompt on the feedback booklet eliciting a response about how the test can be used in a teaching context, the discussion on the pedagogical aspect of test use naturally occurred among the participants in all four groups. This finding seems to reflect the nature of a diagnostic

test. As Alderson (2005) underlines, a diagnostic test should be an interface between learning and assessment. It is easy for test developers to lose sight of the teaching aspect. It seems that those reviewers who were more inclined towards teaching than testing needed more information about the role of the test in an educational context and how it could help students and teachers.

Not entirely unrelated to this category of learning related questions were those pertinent to how feedback could be provided to the learners based on the test, for example:

> How would you give feedback?
>
> What would you do if you have a class with wildly varying scores?
>
> How do you devise a suitable course of action?
>
> Is it just the students or teachers devising the action plan based on the results?

Pedagogical feedback based on test is not usually discussed in the specs document. For example Davidson and Lynch (2002) do not explicitly list feedback based on the test as a part of what should be included in a spec. Also, pedagogical feedback may not be discussed elsewhere in the test development documents or manuals. Nevertheless, in the current study partly because of the nature of the test and partly because reviewers were language teachers, they needed detailed information about how feedback is provided.

More importantly, the type and frequency of questions raised by reviewers about the use of the test in a learning setting are interesting in the context of AFL where it is argued that testing and teaching considerations must be bridged (Assessment Reform Group, 2002; Black, Harrison, Lee, Marshall, & William, 2004; William, 2011). One way this bridging can take place dynamically is when both teaching and testing considerations are discussed at the initial stages of test development. As evidenced above, reviewing test specifications seems to provide a productive platform for teachers and testers to exchange ideas and revise the blueprint of the test such that it organically embeds features essential for achieving intended effects of learning at a specific instructional context. For example, the reviewers of this study suggested that the specs should be revised so that scoring segment becomes more explicit and accessible to test users including teachers and learners in order that they could provide a more straightforward feedback and clearly see the link between the score and the profile of learners' performance on the test in detail. Moreover, there were some concerns voiced about the possible negative washback effects of certain items.

It should be noted that the ultimate goal of a specs review is to build a stronger validity argument by identifying issues that might undermine the validity. In the test development process there are instances where validity can be threatened, for example, if the specs are not accessible to item writers, they may not refer to it and follow their own agenda in writing the test items (see Kim, et al., 2010).There are also instances where expertise of reviewers can be utilized in identifying potential problems with the test even before pilots and tryouts to save resources: for example, in this study consequences of the test and the changes that had to be made to the test to make it more useful from a pedagogical point of view (a.k.a Messick's (1989) consequential validity) were hotly discussed by the reviewers from the early stages of the test development project.

The current study found indications that to improve the specs of a test in the context of AFL where teachers are actively involved in interpreting the test results and modifying the teaching and learning activities (Black et al., 2004), the specs review group could include mixed ability experts i.e., teachers, testers, NSs and NNs. Including teachers in the review team can help with embedding

learning consideration into the specs. This helps the specs become more aligned with the expectations of teachers and the specific requirements of learning situation. Thus, such a review can be a step in the general direction of improving a priori validity of a test.

The second research question was about the percentage of items that are suggested to be revised by the reviewers and item analysis. Spec review participants, after reading the specs and sample items, suggested that the following 17 items needed to be revised: E11, 12; and MC 1, 8, 10, 12, 14, 15, 18, 19, 20, 24, 15, 19, 30, 31, 32. This qualitative appraisal of the items was compared to the results of a quantitative analysis.

In the quantitative item analysis phase, the purpose was to detect potential problems with items and quantitatively verify some of the concerns raised in the review workshop about the 17 items considered as poor by the spec reviewers. To this end, the following steps were taken in analyzing all the items on the test:

Item Facility: According to Brown (2012)CTT prefers items of intermediate facility index .30 to .70. However, to increase the diagnostic power of the test to detect variability in the knowledge of the so called weak students, only items that had an IF above .85 were selected to be revised: Only one such item was found i.e., E1.

Item discrimination: In CTT items that enjoy higher r.pb will contribute to the reliability of the whole test (Brown, 2012). In the analysis, the items with r.pb below .3 were reviewed closely to decide whether they needed to be revised. Six such items were found: (E1, E15, E18, MC24, MC30, MC32) one of them had a typo error i.e., MC 32; and one had very high IF i.e., E1. However, some of these items had r.pbs that were only marginally below .3 and showed particularly good IFs. An analysis of their content indicated that they possessed diagnostic value and could provide teachable moments.

Distractor analysis: Only the part of the test with MC format was analyzed for distractors. Problems were identified with non-functioning distractors in the following items, as one or more of the distractors on these items had attracted less than 5 percent of the test-takers: MC1, 2, 3, 5, 8, 10, 11, 14, 15, 17, 25, 27, 28, 31. The reason for having a low criterion for attractiveness for the MC items was that the feedback for the test is individualized and it might provide insights to the mindset of the test-taker who chooses a distractor that is not chosen by many others. Analysis for mal-functioning distractors was not conducted because a wrong choice by a high level student produces a teachable moment provided that the test is a low-stakes diagnostic one in which spread of the scores for selection purposes is not a priority.

Considerations for improving the specs. Only one pattern was observed in the items that were ultimately revised. The majority of them were measuring tense and aspect; more specifically those aimed at measuring the knowledge of negation and order of parts of speech in a sentence. The systematic problem was lack of functioning distractors. The problem of negation items was dealt with by using students' common errors, as experienced by teachers such as Jane. The problem with the distractors of the items that measured knowledge of order of parts of speech was dealt with by making the distractors longer.

Table 2

The Items that were Detected by Reviewers, Item Analysis, and Both

Problematic items	Suggested by the reviewers	Suggested by item analysis	Suggested by both	Items that were finally revised
E1		X		X
E3		X		X
E6		X		
E8		X		
E11	X			
E12	X			
E19		X		X
E22		X		X
E23		X		X
E25		X		X
MC1	X	X	X	X
MC2		X		
MC3		X		X
MC5		X		X
MC8	X	X	X	X
MC9		X		
MC10	X	X	X	X
MC11		X		
MC12	X	X	X	
MC13		X		
MC14	X	X	X	
MC15	X	X	X	X
MC17		X		
MC18	X	X	X	
MC19	X	X	X	

MC20	X	X	X	
MC24	X	X	X	
MC25		X		X
MC27		X		
MC28		X		
MC29		X		
MC30	X	X	X	
MC31		X		X
MC32	X	X	X	X
Total	14	32	12	15

Table 2 shows the items that were suggested for revision by reviewers and item analysis. According to table 2, 34 items were suggested for revision either by the reviewers or item analysis: 14 were suggested by the reviewers, 32 by item analysis, 12 by both reviewers and item analysis. Of these items only 15 items were finally revised to be included in the revised version of the diagnostic test. The reason for not revising the remaining items was twofold: some of the items that were suggested by the reviewers for revision showed particularly good item statistics, and some of those suggested for revision by items analysis were considered to contain valuable content and could provide teachable moments in an AFL context. Thus the final decision whether to revise the items was context dependent i.e., in a different testing context one may need to revise all the items.

In sum, about 38% of the items that were suggested to be revised by item analysis were the same items also suggested for revision by the reviewers. Although this finding is limited to this study, it suggests a qualitative review can, to some extent, identify poor items.

Conclusion and implications

As this paper implies, improving specs is a major undertaking and questions may arise about it being a worthwhile endeavor. As one of the indispensable requirements for increasing an overall quality of a test, it definitely takes an investment of labor to make a set of specs strong; however, the process becomes quicker once the specs writer gains experience and at a given institution certain styles and archetypes will emerge, which also aids in efficacy (F. Davidson, personal communication, July 8, 2012).

Furthermore, relying on teachers in developing tests is becoming an unavoidable reality in many corners of the world. Gardner (2010) argues factors such as rising costs of external testing and dubious contribution of external testing to effective learning has resulted in growing interest in assessment by teachers. However, teachers cannot be expected to develop quality tests without having adequate knowledge of language testing. Moreover, as the bulk of the literature on language testing deals with large-scale assessment, acquiring that body of knowledge may not be the best

investment for a teacher. However, there are elements of test development that are feasible for a classroom setting. One practical solution is to concentrate most of the quality control effort (Saville, 2012) at the development stage by writing specs. This can spark discussions about validity issues right from the genesis of the test and since a posteriori validation is a luxury that the majority of the classroom teachers cannot afford, a priori validation remains as one of the few opportunities for good practice (Davidson, 2012).

Not only for language classrooms but also for many small-scale testing projects where complex psychometric analyses and formation of a variety of specialized task forces are an unaffordable luxury, and where there is a mounting need for accountability, the test developers should find more creative ways to ensure the quality of their products (Davidson, 2008). A spec-driven a priori validation approach seems to be a practical procedure for this purpose. However, as the findings of this study showed a qualitative review of items can only identify some of the faulty ones. Therefore resource permitting, such a procedure can be complemented with more quantitative methods of item analysis and test validation.

Further Research

Research on the process of specs reviewing as one line of validity argument is in its infancy with numerous novel research questions. However, qualitative studies seem to be more essential at this stage as we are still traversing uncharted waters. As an instance, whether a spec has a primary author or is written by a group may have bearings on how one could interpret the findings of this study. The specs for this study had a primary author, thus the findings are more likely to generalize to other settings where test specs are mainly developed by one person and then feedback is sought e.g., in small institutional, or classroom settings. In other settings, i.e., large testing companies and agencies, specs may be written by a group of employees where there might be no primary author to take the lead and be credited and simultaneously stand accountable for its ownership (see Davidson & Lynch (2002) for a discussion of ownership). The authorship/ownership chemistry and how the specs writers become accustomed to each other so that they work in harmony as they all try to know how each of them writes and what is generally expected from them is worthy of future research.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment.* London: Continuum.

Assessment Reform Group (2002).*Assessment for learning: 10 principles.* Retrieved November 2012 from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf. Cambridge, UK: University of Cambridge. School of Education.

Bachman, L. F., & Palmer, A. S. (1996/2010). *Language testing in practice*. Oxford: Oxford University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2004). *Working inside the black box: Assessment for learning in the classroom. Phi Delta Kappan, 86*(1), 8–21.

Chapelle, C. A. (2012). Conceptions of validity. In G, Fulcher and F, Davidson (Eds.), *Routledge handbook of language testing* (pp. 21-33). London: Routledge.

Davidson, F. (2008). The straightjacket and the blessing of the canon. *Language Assessment Quarterly*, *5*(3), 267-274.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G, Fulcher and F, Davidson (eds.), *Routledge handbook of language testing* (pp. 197-207). London: Routledge.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Fulcher, G. & Davidson, F. (2007) *Language Testing and Assessment: An Advanced Resource Book*. Oxford: Routledge.

Gardner, J. (2010). Developing teacher assessment: an introduction. In J, Gardner, W. Halen, L., Hayward, G., Stobart, & M., Montgomery (Eds.), *Developing teacher assessment*. Berkshire: Open University Press.

Inbar-Lourie, O. (2010). Language assessment culture. In E., Shohamy and N., Hornberger (Eds.), *Encyclopedia of language and education* (2nd edition) Volume 7: Language testing and assessment: Springer: New York.

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.

Li, J. (2006). *Introducing audit trails to the world of language testing* (Unpublished master's thesis). University of Illinois at Urbana-Champaign, USA.

Messick, S. (1989). Validity. In R. L. Rinn (ed.), *Educational Measurement* (pp. 13-103). New York, NY: Macmillan.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Saville, N. (2012). Quality management in test production and administration. In G, Fulcher and F, Davidson: *Routledge handbook of language testing* (pp. 395-412). London: Routledge.

Turner, C. E. (2012). Classroom assessment. In G, Fulcher and F, Davidson (Eds.): *Routledge handbook of language testing*. London: Routledge.

Weir, C. J. (2005). *Language Testing and Validation: An evidence-based approach.* Houndgrave: Palgrave-Macmillan.

William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3-14.

Acknowledgments

We are deeply indebted to Prof. Fred Davidson for his guidelines about conducting research on test specifications and for his insightful feedback on an earlier draft of this paper. Also, our thanks go to the editors of IJLTR and the anonymous reviewers for their comments.

Hamed Zandi has a Ph.D. in TEFL from the University of Tehran. He is currently an instructor at the Institute for Advanced Studies in Basic Sciences (IASBS). His interests include language testing, test specification theory, item writing, item analysis, assessment for learning in the context of EFL education, and vocabulary studies.

Shiva Kaivanpanah has a Ph.D. in TEFL from the University of Tehran. She is currently an associate professor at the same university. Her research interests include language testing and assessment, dynamic assessment, vocabulary studies, assessing writing, learning strategies, and language pedagogy.

Seyed Mohammad Alavi has a Ph.D. in applied linguistics from Lancaster University. He is currently the dean for research affairs at the Faculty of Foreign Languages and Literatures at the University of Tehran. He is also an associate professor at the University of Tehran. His research interests include language assessment, language test development, item analysis, and teaching methodology.