International Journal of Language Studies
Volume 14, Number 2, April 2020, pp. 135-146

Language assessment: Lessons learnt from the existing literature

Mohammad Ali SALMANI NODOUSHAN, IHCS, Iran[1]

Language testing has witnessed three major trends in the 1990s: theoretical, methodological, and analytical. Theoretically, emphasis has been placed on the further understanding of the construct of language proficiency. Methodologically, there has been an outburst of interest in language performance testing and the promotion of the professional standards of test development and use. Analytically, emphasis has been placed on the implementation of Item Response Theory (IRT), G-theory, and the understanding of the multiple sources of variance in test performance. After a review of these trends, the current paper presents a complete picture of language assessment from an Archimedean point. It argues that language assessment, seen from such a point, has four intertwined but self-informed pillars: construct issues, psychometrics, edumetrics, and construct-irrelevant factors.

Keywords: Construct Issues; Construct-Irrelevant Factors; Edumetrics; G-Theory; Item Response Theory; Psychometrics

1. Introduction

Although language assessment has always relied heavily on other disciplines (including education, educational psychology, linguistics, language teaching methodology, ESP, statistics, psychometrics, and so forth), it is considered as a subfield of applied linguistics in which a wide range of scholars are working on a diverse range of topics from analytical, methodological, or theoretical perspectives. Such topics pertain to norm-referencing, criterion-referencing, or both (Brown & Salmani Nodoushan, 2015). Norm-referencing includes proficiency, placement, and aptitude testing—albeit with an eye on complex statistical analyses and theories of validity. Criterion-referencing includes classroom or curriculum notions of portfolios, conferences, self- and peer-assessment, task-based assessment, and continuous, differential and dynamic assessments—all of which are part and parcel of diagnostic testing, progress testing, and achievement testing (Brown & Salmani Nodoushan, 2015). What

[1] Affiliation: Institute for Humanities and Cultural Studies, Tehran, Iran

ISSN: 2157-4898; EISSN: 2157-4901

unites language testers is their attempt at applying theories and practices of testing and assessment to languages; what divides them, however, are their different specializations within language testing which can be viewed as a continuum with hard core positivist and/or empiricist approaches at one end, post-modernist interpretive perspectives at the other end, and everything in between (Brown & Salmani Nodoushan, 2015).

This range of diversity in language assessment is so confusing for people not versed in the field that it can cause cognitive tunneling (i.e., cognitive tunnel vision) where too much concentration on a demanding task or topic prevents them from attending to other issues in language assessment/testing or to the overall complete picture of the field. The current paper seeks to go around this cognitive tunneling to present a complete picture of the field from an Archimedean point. I shall first present a brief overview of the theoretical, methodological, and analytical developments of language testing in the 1990s and then describe the four cornerstones of language testing which I see from a falcon's eye perspective on the field.

2. Background

2.1. Language testing before the 1990s

The IATEFL Language Testing Symposium in 1989 was a milestone in the history of language testing which marked the onset of a new trend in research on language testing for the 1990s onward. Prior to the 1990s, language testing would evolve in an incremental fashion triggered by developments in growth and educational psychology which often caused changes in methods of language teaching which also relied heavily on developments in linguistics. Table 1 presents a summary of this evolution (cf., Jafarpur, 1992).

Table 1
The Reliance of Language Testing on Language Teaching, Linguistics, and Psychology Prior to the 1990s

Psychology	Linguistics	Teaching	Testing
Behaviorism	Structuralism	Audiolingualism	Divisibility
Cognitivism	Mentalism	Neo-cognitivism	Indivisibility
Task-oriented	Functionalism	Notional Functional	Pragmatic
Gestalt	—	—	Cloze/C-test

Around the middle of the twentieth century, the 'humanistic' ideas of some psychologists—specifically Maslow (1943, 1967) and Rogers (1969)—caused a great shift in approaches to education which have come to be known as 'humanistic education'. A human being was no longer treated as a monkey in Skinnerian behaviorism—that is, a mindless robot that could be conditioned

to behave in the way he or she was drilled to. Humanistic education, à la Maslow (1943, 1967) and Rogers (1969), was not a new idea; rather, it had its roots running from Socrates through the Renaissance. A human being, from this perspective, is an individual who possesses certain inherent drives that motivate them to move toward self-actualization whereby they can realize and express their own capabilities and creativity. This implies that education should be fair and authentic, and that testing—which follows education—should also be fair and authentic.

It was on this ground that Messick (1988) called the traditional views of test validity into question and opted for an integrative view of test validity that embraced the 'value implications', 'social consequences', 'relevance', and 'utility' of (language) tests. Informed by Kane's (1982) Sampling Model for Validity, Messick (1988) "viewed validity as an integration of complementary forms of convergence and discriminate evidence" (Salmani Nodoushan, 2009, p. 7). This is a unified concept of validity composed of six different but inter-related aspects: (a) the content aspect, (b) the substantive aspect, (c) the structural aspect, (d) the generalizability aspect, (e) the external aspect, and (f) the consequential aspect (Salmani Nodoushan, 2009). Table 2 summarizes Messick's (1988) perspective on validity.

Table 2
Facets of Validity Envisaged by Samuel Messick (1988)

	Interpretation	Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/utility
Consequential Basis	Value Implications	Social Consequences

As Messick (1988) argues, validity is an integrated evaluative judgment of the degree to which "empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *attitudes* based on test scores or other modes of assessment" (p. 13). As such, validity relates to the "evidence available to support test interpretation and potential consequences of test use" (Salmani Nodoushan, 2009, p. 7). This can also include construct irrelevant factors (e.g., washback).

2.2. The 1990s onward

The IATEFL Language Testing Symposium in 1989 was, therefore, informed by this new perspective on validity. It marked the beginning of a new trend in language testing research for the 1990s. It should also be noted that, during the 1980s, there had been a number of large scale testing projects which had been preceded by a 'State of the Art article' written by Davies (1978). These

projects were also based on the works of Alderson and Hughes (1981) and Hughes and Porter (1983). It was not until 1988 when the publication of another State of the Art article by Skehan (1989) reviewed the developments which had been made during the 1980s and suggested those which were to be made in the 1990s. On the whole, the basic trends in the 1990s were going to include the following:

- The nature of language proficiency;
- The effects of test method;
- The influence of personal attributes;
- The appropriateness of Item Response Theory (IRT) models;
- The application of Generalizability Theory (G-theory);
- A wider utilization of Criterion-Referenced measurement techniques;
- A renewed interest in aptitude testing, self-assessment;
- The application of computer technology to language testing.

(Bachman, 1990, pp. 211-220)

According to Bachman (1990), all of these trends can be summarized into two major pathways in the field of language testing: (1) issues relevant to theoretical and practical developments in the general field of language testing, and (2) issues related to assessment procedures, test development projects, and difficulties of implementing innovations.

By the same token, Skehan's (1991) paper picked up themes from his own 1989 state-of-the-art article to draw on the relationship between language testing and applied linguistics. Skehan asserted that the time was ripe for the two disciplines to be cooperative—that is, to learn from each other's (a) developments and (b) methodologies. He then proceeded to present his own personal view of which areas of language testing were likely to be of utmost importance in the 1990s. The major concerns of language testing in the 1990s, à la Skehan, would be:

(a) developing our understanding of the structure of language proficiency;
(b) further exploration of the relationship between models of language competence and the real world of language performance; and
(c) further developments of methods for improving and validating our instruments.

Skehan (1991) favored more collaborative and cumulative research among language testers as the best way for building up a substantial database of knowledge on how to test language ability.

Rapid developments in electronic communications, especially the Internet, made the implementation of such a cumulative database possible, and Bachman's (1990) hierarchical model of Communicative Language Ability (CLA) was considered as a possible framework for such cumulative research. Specifically, the "Method Effect" section of Bachman's model was welcomed by a good number of language testing scholars. Bachman's CLA model was in fact a development from an empirical study of the existing Canale and Swain's (1980) model. Nevertheless, there have also been other innovations in the field of language testing. Two such innovations are: (a) the refinement of the ACTFL/ILR interview, and (b) the development of the TEEP ESP test (O'Loughlin, 2001). These innovations have established the general principles of test design through the process of solving an existing problem.

As such, the major concern in language testing in the 1990s was the establishment of a connection between SLA research, on the one hand, and achievement testing, on the other. It is no surprise then that the major findings of SLA research have greatly influenced the developments in language testing. Back in the 1990s, these findings included:

(a) the notion of variability in interlanguage,
(b) effects of task design, and
(c) interlocutors' effect on the sample for assessment.

Language testing specialist of the time drew on these to propose that three techniques could be identified as promising for the future directions of the filed: (1) group oral testing, (2) series tasks with a storyline, and (3) indirect communicative skills tests marked on required information points (See also Porter, 1991). By the same token, Davies (1991) emphasized the need for language testing to be receptive to developments, changes, and innovations in such diverse areas as second language acquisition. Drawing on the ideas of Alderson (1991) and Pollitt (1991) about reliability and validity, Davies eloquently argued that, in the coming decades, research should concentrate more on the validity side—a point that echoed Messick's (1988) concerns about validity (See Table 2 above).

Later, Douglas (2000, 2001) and Douglas and Chapelle (1993) noted that language testing in the 1990s witnessed (a) theoretical, (b) analytical and (c) methodological innovations and developments. In terms of theory, à la Douglass (2000, 2001), language testing focused on three major issues: (1) more refined models of language ability, (2) a clearer understanding of the nature of reliability and validity, and (3) interfaces between language testing research and second language acquisition research. In terms of methodology, advancements in language testing addressed five important topics: (1) advances in language skills testing, (2) an increased interest in performance

testing, (3) the development of innovative test formats, (4) the development of new test batteries, and (5) an increased awareness of the importance of standards of practice in language testing. Finally, in terms of analysis, language testing developed in three directions: (1) potential new applications of Item Response Theory (IRT) and Generalizability Theory (G-theory) to test performance, (2) the development of analytic tools for criterion-referenced (CR) testing, and (3) research efforts which attempt to investigate the multiple sources of variance in test performance, including the study of test-taking strategies (Douglass, 2000, 2001).

In his theoretical treatment of the "why" of language testing, Davies (1990) offered a 'humanistic account' of the scope and role of language testing in applied linguistics and language teaching. However, Bachman (1990) was more concerned with the "how" of language testing. Taken together, they have laid the 'theoretical' foundations upon which language testing in the 1990s was based—and so is much of the current 21st-century language testing research. However, the 'methodological' foundations of language testing in the 1990s should be credited to Bachman and Palmer (1996), Cohen (1994), Heaton (1990), and Weir (1990, 1993). Finally, 'analytical' developments in (language) testing (mainly, IRT and G-Theory) should be credited to Boldt and Oltman (1993), Choi and Bachman (1992), Hudson (1989), Kunnan (1992), Reynolds, Perkins, and Brutten (1994), Sasaki (1991), and Stansfield and Kenyon (1992). As such, what unites language testing scholars is their interest in applying testing, assessment and measurement to language, but what divides them are their specializations within the field which, à la Brown, range from hard-core advanced and positivist statistical orientations to postmodernist interpretive perspectives and everything in between (Brown & Salmani Nodoushan, 2015).

Such a wide range of expertise within language testing which has its roots in the 1990s has resulted in controversies over the precision of the technical register used in the field. For one thing, the terms 'evaluation', 'measurement', 'testing' and 'assessment' have caused some confusion. To avoid confusion, the term evaluation should be used within such phrases as 'course evaluation' and 'program evaluation' to refer to the processes of determining the (a) value and (b) ways of improving the curriculum of a given language course or program (Brown & Salmani Nodoushan, 2015); Measurement, on the other hand, refers to the ways in which learners' behavior is (a) quantified, (b) coded, or (c) described; this not only includes tests but also questionnaires, observations, and so forth—be they related to language or any other disciplines. Testing (be it numerical or verbal) has to do with the 'summative', 'formative', 'direct' or 'indirect' observation of the language behaviors of language learners; it is done for feedback and decision making purposes (Brown & Salmani Nodoushan, 2015). Finally, assessment encompasses all

forms of testing and measurement and is done to promote learning or for grading purposes. Focusing on 'processes' and 'purposes', assessment seeks to determine the language performance, progress, and achievement of individual students in language teaching and learning situations (Brown & Salmani Nodoushan, 2015). As such, program evaluation encompasses testing and measurement but serves the program; assessment, too, encompasses measurement and testing but serves the classroom.

3. Where are we now?

In his answers to my questions in a friendly interview, James Dean Brown defined language testing/assessment as a subfield of applied linguistics and argued that the scientists and researchers who work in this field approach issues of assessment/testing from a wide range of perspectives which include norm-referencing and criterion-referencing (Brown & Salmani Nodoushan, 2015). He went on to argue that norm referencing includes proficiency, placement, and aptitude testing as well as complex statistical analyses and theories of validity; criterion referencing, à la Brown, includes diagnostic, progress, and achievement testing as well as notions and issues that have to do with classroom and curriculum—for instance, self- and peer-assessments, portfolios, conferences, task-based assessments, continuous assessment, differential assessment, and dynamic assessment (Brown & Salmani Nodoushan, 2015). In other words, Brown seems to summarize the whole field of language testing into two major subfields: (1) criterion-referencing, and (2) norm-referencing; this perspective has undoubtedly informed his monograph *Testing in Language Programs* (Brown, 1996).

4. Discussion

While I do agree with Brown that looking at language testing through a falcon's eyes brings such a picture to your sight, my own perspective on language testing is more informed by the 'theoretical-methodological-analytical' dichotomies—if they can be called dichotomies at all—which were discussed in section 2.2. above; I have my own reservations, though. Much of what we consider as 'language' testing issues is not specific to language testing but to all kinds of testing where human beings and their abilities are to be measured. As such, many of the topics that are collected under the heading 'language testing' (e.g., bias, fairness, test score pollution, ethics, etc.) are in essence not specific to language testing *per se*, but pertain to all instances of testing. More importantly, they are not part of the 'linguistic' construct which we aim to measure—if we can talk about the psychological reality of such a construct in relation to foreign language learners at all.[1]

Around the turn of the millennium, I was a PhD candidate at the University of Tehran where I had my Language Assessment course with Professor Hossein Farhady, the prominent figure in the field in Iran. As part of the course requirements, I broached the present model for language testing/assessment from the existing literature—and in the form of a term paper. Nevertheless, Professor Farhady had his own perspective on language testing which he later refined, expatiated upon, and presented at the *22nd Annual Language Testing Research Colloquium* in Vancouver, Canada (Farhady, 2000). Later, he refined his model again and published it as a chapter in his excellent monograph (Farhady, 2006). Although his model is a considerable one, my perspective on language testing deviates from it in certain ways.

I believe language testing (or assessment), seen from an Archimedean point, has four intertwined but self-informed compartments: (1) edumetrics, (2) psychometrics, (3) construct issues, and (4) construct-irrelevant factors. While my perspective is informed by the totality of the existing literature on (language) testing/assessment, it has its own specific implications and perhaps entailments. A brief description of these compartments is useful as it will show us a complete picture of the field of language testing/assessment.

I take 'edumetrics' to refer to any aspect of (language) assessment (be it related to 'definition', 'measurement', or 'utilization') that is concerned with task authenticity/genuineness—and that acknowledges cognitive complexity on an 'individualistic' basis. As such, my perspective on edumetrics would include any aspect of (language) assessment that is criterion-referenced, and that compares any given individual's task performance with a criterion. Note that this can also apply to other forms of assessment/testing, not just 'language' testing. Psychometrics, on the other hand, is 'norm-referencing' in essence in that it compares learners with each other; it emphasizes the differences between learners on the normal probability curve. Note again that this, too, can apply to other forms of assessment/testing, not just 'language' testing. As such, norm-referencing and criterion-referencing are borrowed into language testing/assessment but are not innate to it; rather, they are part and parcel of testing/assessment in general—where human beings are to be tested/assessed. As such, JD Brown's view of the field of language testing comprises part of my falcon's eye perspective on the field. Nevertheless, my picture also includes two more areas.

Unlike psychometrics and edumetrics, construct issues have to do with how we define the construct to be measured (e.g., proficiency, achievement, ability, etc.); it should be noted that formative and summative assessment are methodological issues that fit in here. Finally, construct-irrelevant factors are factors (like washback, motivation, bias, personality traits, etc.) that are not

part of the construct which we aim to measure, but that affect how precisely we measure it.

5. Conclusion

Prior to the 21st century, models of language testing were mainly based on developments in psychology and language teaching methodology. Any language test that showed adequate traditional reliability, validity, and practicality—i.e., the 'sine qua non' of language testing à la Harris (1969)—would be considered appropriate. Language testing in the 1990s and the 21st century has nonetheless developed in theoretical, methodological, and analytical pathways, and the focus in theory development during the past couple of decades has mainly been on issues related to validity. However, as JD Brown has rightly suggested (and I do agree with him), the focus in the real world of people who have to actually develop and use language tests should be on practicality—albeit with a dash of reliability (and sometimes validity) thrown in for good measure (Brown & Salmani Nodoushan, 2015).

Notes

1. If we agree with the LPS and LLS dichotomy (cf., McLaughlin, 1987), any language ability that comes from the engagement of LPS cannot be a linguistic construct; rather, it must be considered as a psychological construct.

Acknowledgments

I would like to thank Dr Reza Mobashshernia for his careful reading of an earlier draft of this paper.

The Author

Mohammad Ali Salmani Nodoushan (Email: m.nodoushan@ihcs.ac.ir) is Associate Professor of Applied Linguistics/TESOL at Institute for Humanities and Cultural Studies, Tehran, Iran. His main areas of interest are politeness and pragmatics, and he has also conducted a number of studies on language education and assessment.

References

Alderson, J. C. (1991). Language testing in the 1990s: how far have we come? How much farther have we to go? In S. Anivan (Ed.), *Current developments in language testing* (pp. 1-26). Singapore: Regional Language Centre.

Alderson, J. C., & Hughes A. (Eds.). (1981). *Issues in language testing*. London: British Council.

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

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

Boldt, R. F., & Oltman, P. K. (1993). *Multimethod construct validation of the Test of Spoken English* (TOEFL Research Report RR-46). Princeton, NJ: Educational Testing Service.

Brown, J. D. (1996). *Testing in language programs*. New York: Prentice Hall.

Brown, J. D., & Salmani Nodoushan, M. A. (2015). Language testing: The state of the art (An online interview with James Dean Brown). *International Journal of Language Studies, 9*(4), 133-143.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 6*, 67-84.

Choi, I.-C., & Bachman, F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing*, *9*, 51-78.

Cohen, A. (1994). *Assessing language ability in the classroom*. Boston, MA: Heinle and Heinle.

Davies, A. (1978). Language testing. *Language Teaching, 11*, 145-159.

Davies, A. (1990). *Principles of language testing*. London: Basil Blackwell.

Davies, A. (1991). Language testing in the 1990s. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 136–149). London: Macmillan.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing research: A conceptual gap. *Second Language Research, 17*(4), 442-456.

Douglas, D., & Chapelle, C. (Eds.). (1993). *A new decade of language testing research: Collaboration and cooperation*. Washington, DC: TESOL.

Farhady, H. (2000, Mar. 10). Language assessment: Inter, intra, and supra disciplinary interfaces. *22nd Annual Language Testing Research Colloquium*. Paper presented at the LTRC 2000, Hotel Vancouver, British Columbia, Canada

Farhady, H. (2006). T*wenty-five years of living with applied linguistics: Collection of articles*. Tehran: Rahnama Publication.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Heaton, J. B. (1990). *Classroom testing*. London: Longman Group UK Limited.

Hudson, T. (1989). Mastery decisions in program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 259-269). Cambridge: Cambridge University Press.

Hughes, A., & Porter, D. (Eds.). (1983). *Current developments in language testing*. London: Academic Press.

Jafarpur, A. (1992). *A course in language testing*. Tehran: Payame Noor University Press.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, factor and cluster analyses. *Language Testing, 9*, 30-49.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370–396.

Maslow, A. H. (1967). A theory of metamotivation: The biological rooting of the value-life. *Journal of Humanistic Psychology, 7*(2), 93–126.

McLaughlin, B. (1987). *Theories of second-language learning*. London: Edward Arnold.

Messick, S. (1988). Validity. In L. R. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: American Council on Education/McMillan.

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.

Pollitt, A. (1991). Giving students a sporting chance: Assessment by counting and judging. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 60-70). London: Macmillan.

Porter, D. (1991). Affective factors in language testing. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 32-40). London: Macmillan.

Reynolds, T, Perkins, K., & Brutten, S. (1994). A comparative item analysis study of a language testing instrument. *Language Testing, 11*(1), 1-13.

Rogers, C. R. (1969). *Freedom to learn*. Columbus, OH: Charles E. Merrill.

Salmani Nodoushan, M. A. (2009). Measurement theory in language testing: Past traditions and current trends. *Journal on Educational Psychology, 3*(2), 1-12.

Sasaki, M. (1991). *Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling*. (Unpublished Doctoral Dissertation). University of California, Los Angeles.

Skehan, P. (1989). Language testing (Part I & Part II). *Language Teaching, 22*, 1-13.

Skehan, P. (1991). Progress in language testing: The 1990s. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 3-21). London: Macmillan.

Stansfield, C., & Kenyon, D. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal, 72*(2), 129-41.

Weir, C. (1990). *Communicative language testing*. New York: Prentice Hall.

Weir, C. (1993). *Understanding and developing language tests*. New York: Prentice Hall.