ðə dʒərnəl əv læŋgwəʤ tiʧɪŋ ænd ləmɪŋ

J L T L

dɛdəketəd tu haj kwələti risərtʃ ɪn əplayd lɪŋgwɪstɪks

http://www.jltl.org/jltl/

The Journal of Language Teaching and Learning, 2013–2, 58-64

Issues and Remedies in Composite Scoring: A Case of Joint EGP-ESP Test

Razieh Rabbani Yekta*

Abstract

In comparing the high- stakes ESP tests administered in recent years in Iran for Master Degree and PhD levels admission purpose, a vast area of uncertainty arises as to the nature of scoring system used at the large scale for these competitive exams. Where we had the modular PhD entrance exam with a prerequisite EGP Module followed by a Specific Purpose Module, the participants of our Master Degree counterpart test sat for a joint or blended EGP-ESP subtest which then was scored and reported in a composite percentile and interpreted along with other knowledge subtests against a common national norm. In this article, we attempted to address the issues associated with these models of scoring and the problems which are due from it for our accountability system.

Keywords: Composite percentile, joint EGP-ESP tests, scoring system, norm- referenced models

1. Introduction

In Iran's English for Specific Purpose (ESP) context, ESP tests are administered at two high- stakes contexts, one as the Master's Degree Entrance Exam and the second for the admission of the graduates into PhD programs. In this latter case, test was traditionally administered in two stages: 1) English for General Purpose (EGP) as the prerequisite for the second stage which was 2) the Specialized module. These two modules were separate from each other in all the phases of design, administration, scoring and reporting. Something that makes the Master's Degree counterpart worth investigating and at the same time challenging is the joint nature of these two components at all of the above mentioned phases.

This test is a battery of about 8 different subtests (including one English subtest plus four to seven knowledge subtests of different content courses) where the raw scores of individual subtests are averaged to form an overall test battery mean or a composite score. This resultant composite is interpreted with the consideration of a common norm group for all the underlying subtests and examinees' general ability is reported in terms of their weighted mean. Validity of the battery composite is also enhanced by weighting

* Payam- e- Noor University, Isfahan, Iran. E-mail: r_ryekta@yahoo.com

some of the subtests heavily than others in a total score, so that a low score in one subtest with a light weighting can be compensated in the total score by a same raw score obtained from a subtest with a heavier weighting.

As to the language subtest, after experiencing several years of upgrading, our test developers in Iran have finally reached a fixed framework for the language subtest. For almost all of the fields, this subtest consists of two parts: general English and specialized English. The first part starts with 10 vocabulary test items and continues with a cloze test of grammar with a text of non specialized content. The specialized part is composed of three field specific reading testlets, each with about 5 multiple choice (dichotomously scored) items.

In terms of scoring, also, the only number reported for EGP-ESP component of the battery is a percentage of the participants' correct answers to the items which is calculated by converting the summed score obtained from both EGP and ESP sub-components overall to percentage. But the number which is looked at in decision making phase is not this summed score; in the Report Form there are two other scores which are the base of decision making:1) a total composite score (the whole number average of all subtests), and 2) a composite percentile (the percentile corresponding to the total composite score).

After this initial introduction, a question may arise: bearing this in mind that in PhD admission test, each of the general and specific purpose components has its own portion of the stake in the total stakes of the language test, can summing up the scores obtained from each of these components in M.A counterpart and reporting the outcome in the form of a single raw score be meaningful and usable for inter-individual and intra-individual comparison?

Furthermore, such a scoring system may confront the practitioners, decision makers and test developers with several crucial challenges to deal with. What conditions must be met in such a complex test for the test user to obtain a profile of meaningful, interpretable and useful scores for General and Specific components? And what instruments and models are available to the practitioner for upgrading the test in a way that along maintaining its multi-purposeness, the aforementioned limitations in different phases of test design and test use would be removed?

2. Theoretical Background of the Study

Although it is correct that reporting a composite percentile based on a total score provides the possibility of comparing individual's achievement over a broad curriculum areas (Arce-Ferrer, 2010), such a reporting system will be problematic when there are some narrow objective (ibid) involved in subtest design phase of the test (as in the case of the present language subtest with the joint nature of General and Specific). Some areas of uncertainty which are needed to be dealt with in such complex contexts are discussed below. For the purpose of clarity, problems are addressed separately in the order they may be encountered in the test purpose, test architecture and score interpretation phases.

2.1 Purpose, Test Use and Validity

According to Cronbach (1984), "no test can put all desirable qualities into one test". Relevance of this saying to the subject of the present article is described below in terms of the test purpose and test use and further elaborated under the next heading in terms of the test design.

Design and Validity Chaos

Previous investigations (see for example: Torre &Patz, 2002; Johnson & Carlson, 1994) showed that the estimation method for the correlated abilities yields more efficient results when it is based on the simple structure than composite structure items (as in the case of the English component of M.A Entrance Exams held for non-English Majors in Iran); that is, if we have independent item clusters or independent tests for each trait. From test use and validity perspectives, also, blended tests which are multi-purpose or have no closely specified purpose bring about the serious problems of *design chaos and validity chaos* respectively (Chalhoub-Deville & Fulcher, 2003; Fulcher and Davidson, 2009). The reason for the occurrence of these chaotic situations is that, as described by Fulcher and Davidson (2009), in such contexts, we cannot "collect validity evidence" or "create validity arguments" in support of a particular score interpretation.

In restating this validity-related issue in the context of the joint EGP-ESP test, we can say that unless test users don't know the intention behind the composite scoring and cannot link it to the complex architecture of the test, they cannot interpret the scores and infer anything about the status of examinees' General and Specific Purpose language ability singly.

2.2 Test Architecture: Small-size Sub-scales and Dimensionality

The main issue here is that in the case of the concerned test in this article, while the test measures more than one ability, test developers are constrained to have only a few items for each component. Such short lengths may be necessary in a practical sense, but undermine reliable measurement at the part-score (diagnostic scores, skill scores, or objective scores) level, for example for general English and specialized English separately.

Fulcher and Davidson (2009) in their article on test architecture argue that the test assembly can be changed much more readily than other sub-layers of the test architecture. That is, through introducing more item types or removing them, test designers can better represent a domain. Now, the question is that whether such a modification can be done as easily as described by Fulcher and Davidson in our multiple-purpose test of EGP-ESP which is constrained to have only 15 items for each component. Of course there is controversy regarding the constraints on the number of items Natalie Kohlman (2006), for example, argues that norm-referenced tests usually have only one or two items per objective or standard, while Chase (1999) claims that there need to be at least ten items on any one objective (such as an aspect of grammar structure) in order for a test to cover a broad array of standards.

In psychometric terms, also, as Thissen & Edwards (2005) pointed out, many entire assessments are so nearly unidimensional that sub-scores on any parts are simply less reliable realizations of scores on the whole, and as such have little diagnostic value. This is while Zhang and Stout (1999) argue that before anything is reported as to the status of the examinees, something must be known about the true dimensional structure of the test. Otherwise, reporting a single score for a test which is multidimensional, could contaminate what the tests are measuring, how accurately they are being assessed, and how test data are being used in informing the decision making process.

2.3 Score Interpretation

Although issues discussed so far on test purpose and test design phase of the joint EGP-ESP test have some bearing on score interpretation phase (and some of the interpretability problems have already been mentioned briefly in relation to validity and reliability), the main concern which is mostly of very nature of Norm-Criterion referenced debate has been discussed in the remained parts of the article.

2.3.1 A Mix of Domain and System Referenced Components

Under this heading, the joint nature of test and the effects it has on the mode of test score interpretation are discussed. As described before, the first half of the present test comprised of General English test tasks of a strictly system-referenced nature (Baker, 1990; Robinson & Ross, 1996); it means that the test has been "designed to evaluate language mastery as a psychological construct without specific reference to any particular use of it" (ibid); the second half is a Specific Purpose component in which the purpose, the test content and the method are narrowly defined (Tratnik, 2008) and "items must be narrowly and intensively sampled from a specific domain of second language knowledge" (Robinson & Ross, 1996). Regarding the fact that, for the present, only normative scores were reported for this test, is it meaningful to interpret the scores obtained from these two heterogeneous components along the same scale, i.e. in a relative norm referenced term? Or commensurate with the joint nature of test, is it better for the test users to have a mix of relative/absolute score interpretation (ibid) which better fit the components?

Dealing with a similar case, Texas standardized testing system applied a statistical technique of test equating to its norm referenced standardized tests by which a criterion referenced test was administered in a second session and equated with the nationally norm referenced test (Williams, 1989) and in this way both criterion referenced and equated norm referenced scores were obtained. But this model of test scoring had involved some serious flaws of which the problem of norm invalidity (Yen, Green, and Burket, 1987; as cited by Williams, 1989) has been discussed below in the context of the present joint test of ESP-EGP.

2.3.2 Adequacy of the Norms

According to Glutting (2002), one of the important criteria for the normative scores to be meaningful is the adequacy of the norms. On his handout on norm referenced score interpretation, Glutting has referred the readers to The American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council for Measurement in Education (NCME) (1985) for their norm categorization. There are four types of norms: National Norms, Special Group Norms, State Norms and Local Norms of which the first two categories are of relevance to the subject of the present article.

National Norms: the average score on a test (or on several tests) that was achieved by students nation-wide in a specific grade at a specific point in time (retrieved from Glossary of Education, available online at Education.com).

Special Group Norms: When a test is administered for a special group of examinees, who have a special professional or background knowledge, "a better decision can be made on norms based on a pool of the members from those special groups alone because we get to see how each examinee compares to the typical member of that special group" (Glutting, 2002).

Now that in the concerned test of EGP-ESP, examinees' general purpose English ability is going to be tested along their subject Specific English ability, can a single norm based on the general population make the distinction among different groups of examinees with different background knowledge? For example, is it right to rank an examinee's overall English ability above average because his / her summed score in language component of the battery was compared with a special norm group that was below average to begin with! Even worse is when such a joint test is blended with other knowledge subtests in a battery and the whole battery composite score was compared against a common norm group.

3. Suggested Remedies

One general solution to all these challenges is decomposing the entire ***Co***s we have been confronted with in the concerned test including the ***Co***mplex structure items, ***Co***mplex test and ***Co***mposite scores to link the part-scores explicitly to the item clusters, hence to their intended meaning. As to the first two ***Co***s, so far, numerous empirical investigations have been conducted seeking ways to detect the dimensional structure of the complex tests like this (see for example: Boughton, Yao & Lewis, 2006; Yao & Boughton, 2005). Almost all of these studies need highly complex statistical packages. But, quite recently, Rabbani Yekta (2012, the unpublished dissertation) attempted to do the similar dimensionality detection by recoursing to a panel of expert judges in the field of language teaching and testing who became actively involved in reverse engineering of the item specifications (Spec) for the concerned EGP-ESP test. They reached to the consensus that combined with the statistical method, Spec based control of the item assembly in the way that we can have independent clustering and multi-component mapping together in one test (Thissen & Edwards, 2005), can yield more useful part-scores for each component which have a closer match with the dimensions of the test.

As to the last ***Co***s (***Co***mposite scores), there are so many statistical and non-statistical techniques for increasing the usability of part- scores from high- stakes achievement tests (see for example: Arce-Ferrer, 2010). Differential weighting of item clusters was, for example, reported as a way for removing the unreliability in complex tests (Wood, Nye, & Saucier, 2010). But, because some degree of subjectivity is involved in differentiating the weight, this method was not recommended.

In some other methods (Yen, 1987; Weiner et al., 2001; Thissen & Edwards, 2005; Shin, 2004) part-scores were augmented statistically after the item clusters were modified in a way similar to Author (2012, the unpublished dissertation)'s method. In this way part-scores could take some added-valueness over the total score and therefore became interpretable and worth-reporting.

4. Conclusion

The present study is addressing one of the practical issues arising in the interpretation and use of the high-stakes tests when they serve multiple purposes, in this case, as both specific purpose and general English language ability metrics. The remedies proposed here will make a significant contribution to the attempts made to elicit more informative and interpretable data out of such joint tests. Author, in this paper hope to find a way out of the dilemma of validity chaos we've experienced in the history of entrance exams in Iran (see Farhadi and Hedayati, 2009; Farhadi, 1998). The remedies will also have direct implications for the Iranian engineers of test architecture, those who are responsible for the upgrading of the high-stakes tests, occurring almost every year in Iran. Four research areas are of relevance: 1) part-score augmentation, 2) application of multivariate statistical models to item analyses, 3) integration of specific purpose and General purpose language performance data at the design stage, and 4) part score interpretation.

In short, through removing the psychometric limitations of the part-scores as suggested in this article, the high-stakes tests can yield diagnostic information which in turn, informs instruction.

At the practical level, while promoting accountability, enhancing the interpretability of part-scores can make decision-making possible at both inter and intra individual level (Tate, 2004). At an inter-individual level, decision makers can consider one part-score at a time and compare individuals based on that part-score; at an intra-individual level, decision makers consider one individual at a time and compare part-scores made by that individual. The former is of interest when part-scores are used separately to compare individuals (e.g., to rank individuals based on part-scores); the latter is of interest when part-scores are

used jointly to examine score profiles (e.g., to compare part-scores made by an individual and detect strengths and weaknesses).

References

Arce-Ferrer, A. (2010). Investigating approaches to estimate an individual's strand/objective score profile reliability: A Monte Carlo study. Paper presented at the 2010 Annual Meeting of the American Educational Research Association, Denver, CO.

Baker, D. (1990). A *Guide to Language Testing*. London: Edward Arnold.

Chalhoub-Deville, M. And Fulcher, G. (2003). The oral proficiency interview and the ACTFL Guidelines: A research agenda. *Foreign Language Annals*, 36, (4), 498 - 506.

Boughton, D., Yao, K., & Lewis, D. (2006, April). *A comparison of subscale score augmentationmethods using empirical data.* Paper presented at the meeting of the National Council on Measurement in Education, San Fransisco, CA.

Chase, L. (1999). *Contemporary assessment for educators*. New York: Longman.

Cronbach, L. J. (1984). *Essentials of Psychological Testing*, 4th edn. Harper and Row, New York.

Farhady, H. (1998). A critical review of the English section of the BA and MA University Entrance Examination. In the *Proceedings of the conference on MA tests in Iran* (1998). Ministry of Culture and Higher Education, Center for Educational Evaluation. Tehran, Iran.

Farhady, H. & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, 29, 132–141.

Fulcher, G., & Davidson, F. (2009). Test architecture. Test retrofit. *Language Testing*, 26, (1) 123–144.

Glutting, J. (2002). Glutting 's guide for norm referenced test score interpretation, using a sample psychological report. Retrieved from: http://www.udel.edu/educ/gottfredson/451/Glutting-guide.htm

Johnson. E. G., & Carlson, J. (1994). *The NAEP 1992 Technical Report* (Report No. 23-TR-20). Washington, DC: National Center for Education Statistics.

Kohlman, N. (2006). What teachers need to know, and be able to do, about norm-referenced tests. *ELL Outlook*, available online at: http://www.coursecrafters.com/ELL-Outlook/2006/jul_aug/ELLOutlookITIArticle3.htm

Rabbani Yekta, R. (2012). *Justifying the dimensional structure of General Academic –Specific Academic Purposes English subtests for master's degree entrance examinations: An upgrade test retrofit using Rasch de-modularization technique (*Doctoral Dissertation). Isfahan University, Isfahan, Iran.

Robinson, P. & Ross, S. (1996). The development of task-based assessment in English for Academic Purposes programs. *Applied Linguistics*, 17 (4), 455-476.

Shin, D. (2004). A comparison of methods of estimating objective scores. Ph.D. dissertation, The University of Iowa, United States -- Iowa. Retrieved December 10, 2007, from ProQuest Digital Dissertations database. (Publication No. AAT 139395).

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied measurement in education*, 17(2). 89-112.

Thissen, D., & Edwards, M.C. (2005). Diagnostic scores augmented using multidimensional item response theory. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Torre, J., & Patz, R. (2002) A Multidimensional item response theory approach to simultaneous ability estimation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, April, New Orleans, LA.

Tratnik, A. (2008). Key issues in testing English for Specific Purposes. *Scripta Manent*, 4(1) 3 13.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores: ''Borrowing Strength'' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Williams, Paul L. (1989). Using customized standardized tests. *Practical Assessment, Research & Evaluation*, 1(9).

Wood, D., Nye, C., & Saucier, G. (2010).Identification and measurement of a more

comprehensive set of person-descriptive trait markers from the English lexicon. *Journal of Research in Personality*, 44 (2).

Yao, L., & Boughton, K. A. (2005). A Multidimensional item response modeling approach for improving subscale proficiency estimation in cognitive diagnostic assessments. Paper submitted for publication in APM.

Yen, W. M. (1987, April). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, June, Montreal, Quaebec, Canada.

Yen, W. M., Green, D. R., & Burket, G. R. (1987). Valid Normative Information from Customized Achievement Tests. *Educational Measurement: Issues and Practice*, 6, 7-13.

Zhang, J. & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimentional items. *Psychometrika,* 64, 129-152.