Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 192 (2015) 547 – 553

Procedia
Social and Behavioral Sciences

2nd GLOBAL CONFERENCE on LINGUISTICS and FOREIGN LANGUAGE TEACHING, LINELT-2014, Dubai – United Arab Emirates, December 11 – 13, 2014

An Analysis of Gap Fill Items in Achievement Tests

Halka Capkova[a]* Jarmila Kroupova[a] ,Katerina Young[a]

[a]*University of Economics, Prague*

Abstract

This article revives the discussion over measurements of validity in criterion referenced (CR) tests. It presents how the principles of Classical Testing Theory (CTT), normally associated with norm-referenced tests, were applied to the Business English achievement tests at the University of Economics, Prague, Czech Republic. Firstly, measures of validity in criterion-referenced tests, test purpose, and test specifications are discussed. Next, a 10-item vocabulary gap fill subtest is subjected to a detailed analysis through the use of facility and discrimination indices. Key and distractor analyses of each item are then performed. The insights gained from such analyses are examined in relation to the cyclical test design process of constant review of items so that a high level of standardization is achieved. This paper thus provides teachers with simple tools to build valid language gap fill tests which reflect the criteria of accurate and equitable testing.

Peer-review under responsibility of Academic World Research and Education Center.

Keywords: distractor analysis; achievement test; gap fill;validity; construct; facility index; discrimination index

1. Introduction

1.1. Overview of Testing Approaches

Statistical analyses of reliability and validity of individual items have been mainly associated with norm-referenced testing (NRT). In norm-referenced tests, such analyses are used to ascertain that a test reliably discriminates among test takers. In other words, discrimination is the goal on the basis of which important decisions, often regarding selectivity to educational programs or professions, are made. In criterion-referenced testing (CRT), particularly in classroom-based testing situations, the goal is to show a student's progress or achievement. Researchers agree that the use of statistical analyses for such CRT purposes is futile as (Popham and Husek

* Halka Capkova Tel: +5564636
E-mail address: capkova@vse.cz

Peer-review under responsibility of Academic World Research and Education Center.
doi:10.1016/j.sbspro.2015.06.087

Popham, 1971, p.29) explain: "Criterion-referenced measures are validated primarily in terms of the adequacy with which they represent the criterion. Therefore, content validity approaches are more suited to such tests. A carefully made judgment, based on the test's apparent relevance to the behaviours legitimately inferable from those delimited by the criterion, is the general procedure for validating criterion-referenced measures."

Since many institutions now use a cut-off point in form of a percentage, i.e. often an externally imposed „standard" to establish a student's mastery of the content, and make their pass or fail and other administrative decisions based on this result, test designers have even greater responsibility to develop fair and accurate tests. This article intends to demonstrate that criterion-referenced tests are NOT a dead enterprise, as Fulcher and Svalberg rightfully argue (2013 p.1), but form the bulk of a teacher's and student's everyday educational experience. Teachers can thus empower themselves by employing simple norm-referenced (NR) statistical tools in judging appropriateness of individual items on the tests which they create. In this way, norm-referenced and criterion-referenced testing can complement each other instead of bitterly occupy opposing positions. The analysis of the following exercises is an attempt at such synthesis.

1.2. Validity

Test validity is a measure of how accurately a test score reflects a test-taker's real-life language ability. For a test to be valid, all items have to measure what the test is supposed to measure, i.e. what researchers in the field call a "construct". If a test or a section of a test measures vocabulary skills, as in the case of the example in this paper, then all items in this particular section or test need to be generated around this construct. Thus, as information about the test-taker's ability is collected item by item, the inferences made about the test taker's skills from the scores gained on all of these items are justified. In criterion-referenced (CR) testing, validity is demonstrated in the process of test creation. Items are continuously created, piloted, used in real test, analysed and modified based on a clearly stated purpose and test specifications in an on-going cyclical process.

To exhibit validity of the Department's mid-term or final exam tests, this paper employs the principles of Classical Testing Theory (CTT), which are traditionally associated with norm-referenced tests. While norm-referenced tests are designed to discriminate among students, achievement tests are designed to show that students have learned what they have been taught, there is a need for both at the university level. This unique combination gives rise to a situation where teachers want their students to show what they have learned, yet they also have to "separate" out students who achieve significantly below the expected level from those who pass the test with distinction. The values of facility and discrimination indices as recommended by various researchers differ (See Green and Ebel below), but they all reflect the need for some measure of discrimination between the lower and the upper ridge of students. The purpose of university level language tests thus requires the teachers to strike the right balance. It depends on each particular university context how much they want to discriminate among these two ends of the student spectrum and decide on "discrimination" boundaries for their tests. NR methodology is a tool to make them visible.

1.3. Local Situation, Test Purpose, and Test Specifications

There are 19,000 students enrolled at the University of Economics. The English Department offers nearly 50 different courses, including courses for distance students and the University of the Third Age (aimed at the education and stimulation of mainly retired members of the community - those in their third 'age' of life). Analysed scores were selected from the largest compulsory course in terms of enrolment (2,000 students), English for Business Studies, which is targeted at the B2/C1 level of CEFR and is offered every semester. Because of its size and frequency, computer-based testing (CBT) seems to be the ideal option. Not only does it enable the testing of large numbers of students, it also allows the standardisation of tests and test administration conditions. The decisions made about the presentation and delivery are congruent with the test purpose, i.e. gaining information on the achieved level of Business English skills, knowledge, and abilities of large numbers of students on the basis of which they are awarded credits.

In order to prepare students for computer based testing which requires additional skills such as reading from the screen, practice exercises for home preparation in the e-learning section have been created. These exercises closely

resemble the ones used in the CB tests at the end of their course as "students function best on assessment tests if they are constantly challenged by a variety of item-elicitation and item-response formats" (Cohen, 1994, p.170). The students at the University are nearly "digital natives" nowadays, and computers are a part of their daily lives. Therefore, it comes as no surprise that many of them actually prefer e-tests to traditional paper tests. Cohen (1994, p.245) cites multiple studies which summarize several advantages of CB assessment such as minimized frustration and fatigue, immediate test scores and diagnostic feedback, accurate and consistent evaluation of results, improved record-keeping, and easy access to data for research purposes. The final test of the course is therefore computer-based and consists of several types of exercises: open-cloze questions, multiple-choice tasks, gap fill tasks and proof-reading, which are commonly done in CBT. The tested language skills consist of lexis, grammar, listening and reading comprehension skills.

Out of the current database, only multiple-choice gap fills have been analysed so far, and results have been used to restructure and improve the exercises. The selected vocabulary gap fill subtest in this article has been designed to measure the knowledge of economic terminology and concepts in English and the understanding of the vocabulary within the context of the whole passage. It is representative of situations that test developers can encounter when analysing their tests in more detail for validity purposes.

2. Item Analysis

Originally, thirty gap fill tests included in the current database for the English for Business Studies course were carefully analysed. For the purpose of this article, the gap fill exercise given below was selected as the most representative and useful to demonstrate the results. One of the advantages of doing research at the English department is that the sample group can be rather large due to the great number of students taking mandatory English courses. Thus, a random sample of 30 out of 1,800 test takers was chosen for the purpose of this article. A sample of 30 is generally considered large enough to be representative. Statisticians have discovered that at this number a distribution is most likely to approximate a curve of normal distribution (Lawn and Abramowitz, 2008, p.262).

Example of a gap fill item:

---

a) diminish	f) bond	k) bankruptcy	p) rates
b) subprime	g) under	l) stimulate	q) take
c) incomes	h) deregulation	m) regulation	r) elementary
d) stock	i) target	n) deposit	s) defaulted
e) securitized	j) crunch	o) prime	t) undervalued

Subprime crisis In the 1980s and 1990s, 1… and globalization led to great changes in the financial industry. Financial institutions were permitted to do all kinds of financial business. Then, after the dot-com bubble burst in 2000, the Federal Reserve lowered interest rates to 2… economic activity. Low interest rates encouraged lots of people to 3… out mortgages. Many financial institutions began to offer subprime mortgages to people with low 4… . These mortgages often required no 5… payment and had a special low introductory interest rate. But then the housing bubble burst. Meanwhile, the institutions which issued the subprime mortgages had 6… them in mortgage-backed securities and collateralized debt obligations. Unfortunately, 7… mortgages were mixed with subprime mortgages. When house prices fell in 2007, subprime borrowers 8… . Lenders repossessed their homes, the mortgage-backed securities became worthless, 9… markets crashed, and the banking system had hardly any capital left, leading to a credit 10… .

KEY: 1h – 2l – 3q – 4c – 5n – 6e – 7o – 8s – 9d – 10j

2.1. Key Analysis

Ten key terms in this test set were analysed. The table presented below shows the percentage of students who selected the correct option.

Table 1. shows the percentage of students who selected the correct option

1	2	3	4	5	6	7	8	9	10
22/30	21/30	29/30	26/30	10/30	16/30	12/30	24/30	21/30	28/30
.73	.70	.96	.86	.33	.53	.40	.80	.70	.93
73%	70%	96%	86%	33%	53%	40%	80%	70%	93%

The classical approach to analysing how students perform on each item (item difficulty) and how well each item discriminates the better performers from the poorer ones (discrimination index) was used. Both indices are based on how well a test taker performs on a test as a whole against how s/he performs on a particular item. "A total score on a test represents some aggregate of performance across all test items for a specific ability or domain of knowledge" (Haladyna, 2004, p.4). If a test taker does well on the test as a whole, it is expected that s/he will get only the difficult tasks wrong. If it is discovered that this is not true, there is a chance of a weak discrimination index on those particular items, wrongly chosen key terms, or incorrectly set distractors. Facility index or item difficulty refers to the ratio of correct responses to total responses for a given test item, or whether the item is easy or difficult. Cohen advises that "if the assessment instrument is criterion-referenced, and aims to determine whether nearly all students have achieved the objectives of a course, then the assessor may wish to consistently obtain item difficulties of 90% or better" (1994, p.102).

Ideally, the figure showing item difficulties should be around .5 as this value suggests that the item discriminates well and is neither too difficult nor too easy. From Table 2.1. it is obvious that item 6 (FI .53) fully corresponds to the ideal item difficulty. As Fulcher also explains, "we never have a test populated only from items with a facility index of .5. All tests have a range of items, and the general rule of thumb is that items within a range of .3 to .7 should be included" (Fulcher, 2010, p.182).). Green further attests that "in an achievement test, we might expect to find higher facility values than in proficiency tests. For example, they may be in the 80 to 90 per cent bracket, suggesting that the students have understood what has been taught"(2013, p.27). Alternatively, "Facility values between 20 and 80% (.2 to .8) can also provide useful information provided the items still discriminate and contribute to the test's internal consistency". (Green, 2013, p. 26) For the gap fill test that has been analysed, this means that items 1 (FI .73), 2 (FI .70), 5 (FI .33), 7 (.40), 8 (FI .80) and 9 (FI .70) could also be considered as appropriate.

Consequently items 3 (FI .96), 4 (FI .86) and 10 (FI .93) were reviewed. When looking at these more closely, it is obvious that the awareness of verb noun collocations such as in item 3 take out mortgages, the background knowledge in case of the item 4 incomes, and in the item 10 the adjective noun collocation credit crunch is quite high and students acquired it well. Therefore, the high FI is acceptable and actually welcome due to the fact that in economic English, students are required to memorise and actively use these collocations. They also have to fully understand the whole text rather than individual phrases or sentences and to apply their knowledge of economic theory, which is also taught in the course. All these values can be justified if further analysis establishes that the items discriminate well.

Facility values should always be considered and interpreted together with the discrimination indices as discrimination tells us about the extent to which the items separate the better test takers from the weaker ones. It is based on how well a test taker does on the test as a whole in comparison with how s/he performs on a particular item. It seems obvious that if a test taker does well on the test as a whole, s/he is expected to answer an average item

correctly and only makes mistakes in the difficult ones. If not, it may be caused by a weak discrimination index on the particular item and vice versa. If s/he does poorly on the test as a whole, he/she would probably get average and difficult items incorrect. The reason for contradictory results could be again caused by a weak discrimination index.

"Another way to calculate item discrimination is by means of a point-biserial correlation, which measures an item's reliability. A correlation is made between all respondents' performance on a given item and their performance on some more general criterion, usually score on the test as a whole. The higher the point-biserial correlation for a given item, the more likely that the respondents getting a particular item right are also those who perform best on the total test" (Cohen, 1994, p.103).

In order to calculate discrimination index DI, the test takers are divided into three groups – top scorers, middle scorers and low scorers. Then FI for the best group, FI Top, and for the lowest one, FI Bottom, are calculated. In the second stage of calculating, the previously calculated FI of the lowest scorers (FI Bottom) is subtracted from the results of the highest scorers (FI Top). So that DI = FI Top – FI Bottom. According to Popham and Husek "in an achievement test an unsatisfactory item would be one which could not properly discriminate between the more and less knowledgeable learners (as reflected by total test performance). Non-discriminating items are usually those which are a) too easy, b) too hard, and/or c) ambiguous." (1971, p.30)

Table 2. Non-discriminating items

	1	2	3	4	5	6	7	8	9	10
10	9	8	10	10	6	9	6	10	9	10
FI Top	.90	.80	1	1	.60	.90	.60	1	.90	1
10	5	5	9	7	2	1	3	5	5	8
FI Bottom	.50	.50	.90	.70	.20	.10	.30	.50	.50	.80
DI	.40	.30	.10	.30	.40	.80	.30	.50	.40	.20

Items with a DI of .3 and higher will be retained. According to Green, items with a DI of .25 may also be seen as acceptable. "On an achievement test, the discrimination may be low simply because all the test takers are performing well and therefore the amount of variability between the test takers is low" (2013, p. 29).

(Ebel, 1979, p.267) offers this table regarding levels of discrimination:

1. .40 and above very good items
2. .30 to .39 reasonably good items but possibly subject to improvement
3. .20 to .29 marginal items, usually needing and being subject to improvement
4. .19 and below poor items, to be rejected or improved by revision

This calculation proved again that items 3 (DI .10) and 10 (DI .20) do not offer useful information about the test takers on the item but they do not discriminate well either as they fall into the Ebel's categories 3 and 4. On the other hand DI analysis shows that items 2, 4 and 7 (DI .30) discriminate quite well as they fit within category 2. In this case all distractors should be carefully analysed to reveal whether they distract satisfactorily or should be substituted. "For criterion-referenced tests the use of discrimination indices must be modified. An item which does not discriminate need not be eliminated if it reflects an important attribute of the criterion, such an item should remain in the test." (Popham, Husek 1971, p.30) The discrimination index of the remainder of the items corresponds to category 1. These are items 1 (DI .40), 5 (DI .40), 6 (DI .80), 8 (DI .50) and 9 (DI .40).

Further examination focused on key analysis and consequently distractor analysis. Items in the subtests are not weighted differently, so in gap fill verb-noun collocations, adjective-noun collocations and students' ability to use background knowledge are tested equally in order to keep all tests standardized. In the example the following key terms were tested:

Table 3. key terms

Verb-Noun Collocations	Stimulate Economic Activity
	Take Out Mortgages
Adjective-Noun Collocations	Deposit Payment
	Prime Mortgages

Background Knowledge	Stock Market Credit Crunch Deregulation Incomes Securitized Defaulted

Below are presented the distractors chosen for each key term. The next crucial step was to analyse the frequency of the intended distractors towards their key terms and also to other key terms, as it is necessary to also consider that some key terms could serve as a distractor as well. Green says that "real distractors must have been chosen by at least seven per cent of the test takers" (2013, p.34). In the case of the random sample of 30 students, it means that the real distractor is the one preferred by at least two students. The scores of the intended distractors for the key terms are analysed in the table below.

Table 4. key terms

	KEY TERMS	DISTRACTORS	RESPONSE FREQUENCIES
1)	H) Deregulation	M) Regulation	5 Students
2)	L) Stimulate Economic Activity	A) Diminish	1 Student
3)	Q) Take Out Mortgages	I) Target	0 Students
4)	C) Incomes	P) Rates	2 Students
5)	N) Deposit Payment	G) Under	0 Students
6)	E) Securitized	T) Undervalued	9 Students
7)	O) Prime Mortgages	R) Elementary	6 Students
8)	S) Defaulted	K) Bankruptcy	3 Students
9)	D) Stock Market	F) Bond	0 Students
10)	J) Credit Crunch	B) Subprime	0 Students

From the table 4 it seems clear that distractors created for the gaps 1, 4, 6, 7 and 8 were suggested, chosen and created well. On the other hand for the gaps 2, 3, 5, 9 and 10, the distractors were not chosen correspondingly, which is true up to a certain level as in case of the gap 2, key l) *stimulate*, four students chose distractor r) *elementary* and three chose g) *under*. A similar situation was in gap 5, item n) *deposit*, when even though no one chose the intended distractor which was g) *under*, distractors r) *elementary* and a) *diminish* worked as the real distractor as they were chosen by nine students. Also for the key term d) *stock* in gap 9, when instead of the intended distractor f) *bond* students chose b) *subprime* which should have served as a distractor to j) *crunch*. A similar situation happened in the key term j) *crunch* in gap 10 where two test takers chose k) *bankruptcy* instead of the intended distractor b) *subprime*.

Moreover in the case of two items, key terms themselves served as distractors. In the fifth gap these were the key terms o) *prime*, e) *securitized* and s) *defaulted* chosen by seven students and in the seventh gap the key term e) *securitized* served as a distractor for six test takers.

To summarize the results of the distractor analysis, it is clear that at least the distractors for gaps 3 and 10 were not chosen well and need to be rewritten.

3. Conclusion

Borrowing basic concepts from Classical Testing Theory, the test development team members were interested in investigating whether their exercises on a university-level, language achievement test discriminated between top and low achievers. Therefore, they reviewed a database of 30 gap fill tasks (each consisting of ten items plus twenty distractors) and then performed a detailed analysis of one vocabulary subtest presented in this paper. This analysis presents most of the situations that test developers encounter when analysing tests.

Two key indices were presented: the facility index (FI) and the discrimination index (DI) with explanations of how they are calculated and what they show about the quality of any individual item with a follow up key and distractor analyses. This detailed analysis has clearly identified the items which are not suitable for testing and need to be rewritten.

The analyses used clearly proved that there are some weak items which do not work properly and therefore should be replaced or changed. All aspects taken into account revealed that gaps 3 and 10 do not correspond to the range of both FI, DI and do not serve as the intended or an additional distractor. A similar situation appeared in case of gaps 4 and 9 which showed some satisfactory values but in two other aspects they did not work.

The task ahead for the Department is now to create a complex database of test items. The team will have to modify inappropriately written items and replace the distractors which do not discriminate well according to the test specifications, which should lead to higher standardisation. In this way, the Department test creators can be sure that they test what they say they test and that they do so consistently.

It is the authors' belief that CR testing is still alive in the methodology proposed by this paper. NR statistical tools can be used in congruence with CR methodology of test development. Thus, the authors hope to broaden the scope of tools that test creators can utilise in their design and to inspire other academics at language departments across the region to seek ways of validating their own classroom tests. Analyses from testing settings at other universities in the region would be most welcome to confirm these results.

References

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom*.

Ebel, Robert L. (1979). *Essentials of Educational Measurement,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Fulcher, G. and Svalberg, A. (2013). Limited Aspects of Reality: Frames of Reference in Language Assessment. International Journal of English Studies, 13 (2), 1-19.

Green, R. (2013). *Statistical Analyses for Language Testers*. New York: Palgrave MacMillan.

Haladyna, T. M. (2011). *Developing and validating multiple-choice test items*. New York: Routledge.

Lawn, S., Abramowitz, S. K. (2008). Statistics Using SPSS: An Integrative Approach. Cambridge: Cambridge University Press.

Popham, W. James. (1971). *Implications of Criterion-Referenced Measurement* In Popham, W. J., *Criterion-Referenced Measurement.* Englewood Cliffs, NJ: Educational Technology Publications.