Languages for Special Purposes in a Multilingual, Transcultural World

Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria

http://lsp2013.univie.ac.at/proceedings

Between a rock and a hard place: Test security and validity in LSP testing

Romana Zeilinger; Hans Platzer

Cite as: Zeilinger, R. & Platzer H. (2014). Between a rock and a hard place: Test security and validity in LSP testing. In G. Budin & V. Lušicky (eds.), *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria*. Vienna: University of Vienna, 332-339.

Publication date: July 2014

ISBN: **978-3-200-03674-1**

Between a rock and a hard place: Test security and validity in LSP testing

Romana Zeilinger
Institute for English Business Communication, Vienna University of Economics and Business
Austria

Hans Platzer
Department of English, University of Applied Sciences Wiener Neustadt
Austria

Correspondence to: hans.platzer@fhwn.ac.at

Abstract. This paper describes the validation of an end-of-term Business English exam. In this context, we seek to determine whether the exam has sufficient reliability and if it accesses test takers' background knowledge in business. A sample of 320 test papers was analysed. The results indicate mediocre item discrimination, and reliability is, consequently, only marginally acceptable, with Cronbach's alpha of 0.79. Nonetheless, for a non-piloted test, these are respectable figures. In addition, failing students with results below the 60% cut-point may re-sit the exam if they scored above 55%. This threshold represents the lower bound of the 77% confidence interval ($CI_{.77}$) of the cut-score. This safety margin ensures that we can be reasonably certain that failing test takers whose true score might nonetheless be a passing one get the chance to retake the test. Regarding field-specific background knowledge, the findings of a MANOVA suggest that students who performed better in their business classes also achieved significantly higher scores in the Business English test. Overall, we therefore conclude that it is possible to develop a test instrument which addresses field-specific content knowledge and is sufficiently reliable despite the absence of pre-testing.

Keywords. Business English, English for Specific Purposes, language testing, test reliability, test validity.

1. Introduction

This paper describes the validation study a Business English exam administered by the Institute for English Business Communication at the Vienna University of Economics and Business (WU). One key issue in this type of ESP testing concerns its theoretical grounding, which some studies have judged to be rather tenuous. According to O'Sullivan (2006: 174), for example, most ESP testing is regarded as "'industry-driven' with a more pragmatic than theoretical foundation". This is a view that can be traced back to Davies (2001), whose investigations into earlier versions of IELTS led him to conclude that "there is no theoretical basis for LSP testing, […it] remains of uncertain value and, indeed has not proved itself to be more valid than a general proficiency test" (Davies 2001: 144). (Cf. also Thighe (2007) and Ingham and Thighe (2006) on the role of ESP testing.) In this context, Davies (2001) also refers to Robinson (1989), whose position on ESP he regards as merely pragmatic and pre-theoretical. Despite these misgivings, Davies (2001) ultimately accepts Robinson's (1989) point that ESP "is goal-oriented [and that] students study […it…] because they have to perform a task in English" (Robinson 1989, quoted in Davies 2001: 144), and he regards this argument as sufficiently persuasive to continue to engage in ESP testing.

As if in anticipation of Davies' (2001) criticism, Douglas (2000) had already made an attempt to provide a theoretical underpinning for LSP tests. Based on Bachman's (1990) and Bachman and Palmer's (1996) theories of language testing, Douglas' framework is probably the most comprehensive one to date, and also proved influential in O'Sullivan's (2006) revision of Cambridge ESOL's BEC suite of exams. Both researchers maintain that LSP testing rests on the central concepts of specificity and authenticity. Specificity, in particular, is not regarded as

a case of all or nothing, but as being located on a cline, depending on "the amount of content or background knowledge [required] in responding to the test tasks" (Douglas 2000: 14). Indeed, background knowledge is a core concept in LSP, despite the fact that its role is intriguingly complex and still far from clear. Clapham's (1993, 1996) findings, for instance, suggest that test takers with poor language skills are typically unable to exploit their background knowledge, in contrast to more competent language users. In general, the strongest effects of background knowledge were found in test takers with intermediate language competence, and effects proved to be stronger the more specific the test, i.e. the more background knowledge was required. Interestingly, the impact was weakest among the most competent language learners as these managed to use their language skills to compensate for possible shortcomings in background knowledge.

On the basis of Bachman and Palmer's (1996) concept of the topical knowledge component, Douglas (2000) envisions three ways of integrating background knowledge with language skills: (1) a zero option, which only covers (decontextualised) language ability, (2) specific background knowledge and language ability as separate abilities or (3) specific purpose language ability as the result of the interplay between language ability and topical knowledge. Douglas (2000) argues that where the difference in test taker performance on an LSP test may be due to either a lack of background knowledge or language ability, it may be advisable to keep these two constructs separate (i.e. option 2). Conversely, where a solid basis of similar background knowledge can be assumed, "language and background knowledge would be left intertwined" (Douglas 2000: 20). In the present context, neither the course goals of the relevant Business English class nor the reporting of students' results envision any differentiation between a business and a language component. Instead the aim is to measure a combined language-cum-business construct. Douglas' (2000) third option of an integrated concept covering language skills and business knowledge therefore seems to best reflect the test construct of the class in *English Business Communication 2* (EBC2). This paper, consequently, seeks to demonstrate that business knowledge significantly affects test takers' EBC2 scores, thus confirming that the test addresses the combined business and language construct. This should substantiate our expectation of the criterion-related, concurrent validity of the EBC2 exam.

The second topic addressed in this study concerns the impact of test security on another facet of the EBC2 exam's validity, viz. its scoring validity (Weir 2005), involving a.o. item quality and reliability. Piloting a test is universally regarded as essential in ensuring item quality and reliability (Bachman & Palmer 2010; Hughes 2003). However, such pre-testing requires access to trial test takers who have no contact to the actual candidates, otherwise test security is compromised. Unfortunately, no such test takers are available in the setting at WU, hence piloting the EBC2 test items is not an option. In the absence of such pre-testing, it would theoretically also be possible to validate a test that had already been administered, and re-use parts of the now validated test in a later administration. However, candidates at WU have the right to inspect their test papers, which also involves taking copies of the exams, so that these invariably end up in the public domain. Re-using such tests is clearly not feasible, again due to concerns over test security. This state of affairs means that unpiloted tests have to be administered for security reasons, and while students obviously appreciate the transparency associated with being able to inspect their papers, this is clearly at the expense of validity. In particular, the absence of pre-testing prevents test developers from assessing item quality, which has a direct impact on test reliability. Consequently, a second aim of this paper is to determine whether the EBC2 test has sufficient scoring validity despite the absence of pre-testing.

2. Method

2.1. Setting, subjects, test instrument

The test instrument to be validated is the semester final exam of the class in *English Business*

Communication *2* (EBC2). All WU students enrolled in two different Bachelor's programmes, viz. *Business, Economics and Social Sciences* and *Business Law*, are required to take this exam. It was administered by the Institute for English Business Communication on 26 January 2010 and lasted 90 minutes. 320 tests were investigated. The test is divided into three main parts, viz. Section 1 *Business content and terminology*, Section 2 *Language (grammar and vocabulary)*, Section 3.1 *Text comprehension* and Sections 3.2 and 3.3 *Text production*. Sections 1, 2, and 3.1 consist of selected response items or constrained, constructed response items. Examples 1 - 3 below are representative instances of items from these three test parts:

(1) *Section 1 - Business content and terminology*

Rubric: Responses to cases. Answer the following CLEARLY. You need not write complete sentences.

Item 1.1 (b) You have received an accepted time draft in connection with an export transaction. Name the three options you now have as regards receiving payment.

Key: 1. present at maturity; 2. discount; 3. indorse/endorse

(2) *Section 2 - Language (grammar and vocabulary)*

Rubric: Collocations – Text completions. Complete each of the short texts below by filling the gap it contains with the grammatically-correct form of a verb (possibly including a preposition or adverb) that collocates with the accompanying business term.

Item 2.2 (a) The dollar has depreciated against the euro over the last few months. As a result, direct investment in the US, such as _____ firms in the US from scratch, has become much easier for euro zone companies.

Key: establishing; setting up; founding; starting (up); building (up)

(3) *Section 3.1 - Text comprehension*

Rubric: Mark each of the statements below as TRUE (T) or FALSE (F) according to the text. Then write in the space below it the EXACT WORD(S) FROM THE TEXT that support(s) your answer.

Item 3.1.3 (a) Many governments used fiscal measures to stimulate economic growth.

Key: True "Governments worldwide raised their spending spectacularly."

Items from these three parts, i.e. Sections 1, 2 and 3.1, are validated in the present study. On the other hand, Sections 3.2 and 3.3 are pure constructed response tasks (i.e. free text production) and require a completely different approach in terms of scoring validity. They are therefore not included in the present validation effort.

The parts validated here consist of 46 items (Section 1, 15 items; Section 2, 16 items; Section 3.1, 15 items). A total score of 65 marks can be achieved on these sections, with a cut-score of 60% (i.e. 39 marks). The pass rate on this reduced part of the test was 56.9%, i.e. 182 of 320 test takers. As just mentioned, students need to reach the cut-score of 60% for a passing grade. Failing students who nonetheless achieve a result of between 55% - 59% may resit the exam once. The test is consequently a medium stakes exam as such marginally failing students may re-take the exam while other failing candidates with results lower than 55% can repeat the whole course up to twice.

2.2. Validation procedures

As discussed above, this validation effort is based on the concepts of scoring validity and criterion-related validity. Weir (2005) introduces *scoring validity* as a superordinate term

for a test's "statistical attributes" (Weir 2005: 43), including concepts such as item analysis, reliability and measurement error. Against this background, we report item facility (IF) and item discrimination (item-total correlation), and compute Cronbach's alpha as a measure of internal consistency and the associated standard error of measurement, both of which are especially critical in determining if the test instrument has sufficient scoring validity.

To confirm criterion-related, concurrent validity, we employed a differential groups design. The test takers were divided into two groups according to the extent of their background knowledge in business studies. For this purpose, WU Academic Controlling kindly made available to us students' mean grades (n=318) which were based on eight different classes in business administration, viz. Accounting and Management Control 1 & 2; Procurement, Logistics and Production; Corporate Finance; Marketing; Personnel Management, Leadership and Organisational Behaviour; Business Information Systems 1; Introduction to Business Administration. These grades were used to divide the sample into two percentiles resulting in a high performing group (n=178) and a low performing one (n=140), based on their business knowledge. This grouping variable became the independent variable in a multivariate analysis of variance (MANOVA), which aims to determine the impact of business knowledge on the EBC2 exam. Accordingly, we hypothesise that if the EBC2 exam tests business knowledge as well as language skills, one would expect to see a significant effect size of the independent variable (business knowledge) on the test scores. A significance level of 0.01 is used throughout this study. In addition, we assume that the effect size of business knowledge will be highest in Section 1 as this expressly tests business concepts, viz. *Business content and terminology*. By contrast, effect sizes should be lower in Section 2 (*Language: grammar & vocabulary*) and Section 3.1 (*Text comprehension*). If these points can be confirmed, it should go some way to demonstrating that the EBC2 exam tests a business as well as language construct, which in turn should underscore its concurrent validity.

3. Results and discussion

3.1. Scoring validity

Item facility (IF) is typically the first measure to consider in terms of item performance. It is defined as the percentage of correct answers for each item, and according to Bachman (2004: 138) the rule of thumb in test development is to select "items that fall between a range of [...] .20 and .80", i.e. those that are answered correctly by between 20% and 80% of test takers. (Cf. Brown and Hudson (2002) and Carr (2011) for broadly similar guidelines.) Against this background, item facility, as outlined in Tab. 1, seems largely satisfactory. In each of the parts, the majority of items (between 60.0% and 80.0%) falls within the IF range of 0.50 - 0.79 and another 12.5% to 26.7% of items have IF values between 0.20 and 0.49. That means between 86.7% (Section 3.1) and 93.3% (Section 1) of items fall within the required IF range, which is a promising start to the examination of item quality.

IF	Section 1 (15 items)	Section 2 (16 items)	Section 3.1 (15 items)
0.80 - 1.00	1 (6.7%)	2 (12.5%)	1 (6.7%)
0.50 - 0.79	12 (80.0%)	12 (75.0%)	9 (60.0%)
0.20 - 0.49	2 (13.3%)	2 (12.5%)	4 (26.7%)
0.00 - 0.19	0 (0.0%)	0 (0.0%)	1 (6.7%)

Table 1: Item facility in EBC2 exam

However, item discrimination is a more crucial aspect of item behaviour as it directly affects the reliability of the whole test. In the present instance, we will look at item-total correlation as a measure of item discrimination (see Tab. 2). Bachman (2004: 138) recommends the selection of "items that have discrimination indices equal to or greater than .30". (Brown and Hudson (2002) and Carr (2011) suggest virtually identical coefficients.) Items with discrimination values

below 0.30 are candidates for revision, while items with negative discrimination are particularly problematic. Negative discrimination means that weaker test takers tend to do well on such items, while better performing candidates do not. Such items, consequently, confound the measurement of the underlying construct and should, therefore, be scrapped outright (Hughes 2003).

From this perspective, the items in the EBC2 exam look somewhat more problematic than from the point of view of item facility. As outlined in Tab. 2, just over a third of the items (34.8%) feature an acceptable discrimination index of ≥0.30, while almost two thirds (63.0%), i.e. the majority, discriminate only poorly between strong and weak test takers. This is our first indication that if piloting items were possible in the context of test development and test administration at WU, item quality might be enhanced. Most worryingly, one item has negative discrimination, and would under normal circumstances be deleted or replaced. Overall, we are therefore faced with mediocre to poor item discrimination, and this will invariably have repercussions on reliability estimates.

r (item, total)	Number of items
0.30 - 1.00	16 (34.8%)
0.00 - 0.29	29 (63.0%)
< 0.00	1 (2.2%)

Table 2: Item discrimination in EBC2 exam

Cronbach's alpha is a widely reported measure of reliability and internal consistency. For most purposes, a value of 0.80 is regarded as the minimum expected reliability of a test (Carr 2011, Bachman 2004, Brown & Hudson 2002), although higher values should be achieved by high stakes tests. However, the respective reliabilities also vary with the skills tested and the relevant scoring procedures (Hughes 2003). In the present test, Cronbach's alpha comes in at 0.79, i.e. just under the expected minimum of 0.80. Considering that this is based on a non-piloted test, a reliability of 0.79 is actually quite respectable and indicates the robustness of the test development procedures.

These procedures were in fact fairly elaborate. Five different authors of the course book for the EBC2 class collaborated in producing a first version of the necessary test items. The Quality and Examinations Officer compiled these items into a first test draft and at the same time already made some initial revisions. This revised draft was reviewed by all five test developers, and subsequently the Quality and Examinations Officer produced the final draft incorporating the feedback from the test developers and further revisions of his own. It is clear that these rigorous test development procedures will need to be upheld in order to maintain the current levels of item quality and reliability, at least as long as test piloting remains unfeasible.

Yet, whatever the reliability achieved, estimates such as Cronbach's alpha are in themselves not very intuitive. Their practical value rather lies in the possibility to compute a test's standard error of measurement (SEM), which is based on a given reliability estimate (such as Cronbach's alpha) and the test's standard deviation (SD), hence $SEM = SD\sqrt{(1-\alpha)}$ (Bachman 2004: 172). On this basis, the SEM (of the three validated test sections) is 4 marks, i.e. we can be 68% certain that a test taker's true score falls within ±4 marks (or 1 SEM) of their observed test score. But what makes the SEM particularly valuable for present purposes is that it allows us to evaluate how reliable our decisions are in failing students who are below the cut-score of 60%. However, before discussing this issue, a closer look at the actual test scores is in order.

According to the test results outlined in Tab. 3, 182 students achieved a passing score (on the three sections discussed here) of at least 39 marks (or 60%). Those that failed can be classified in the following way: 37 candidates, i.e. just over a quarter of fails (26.8%), achieved a score of between 55% and 59%. Regulations at the Institute for English Business Communication stipulate that failing students in this score range may resit the exam. Given a SEM of 4 marks, a test score of 55% happens to be the lower bound of the 77% confidence interval ($CI_{.77}$) of the cut-score. That means if one allows failing test takers with scores between 55% and 59% (36

- 38.5 points) to resit the exam, one can be 77% confident of not making a student repeat the whole course whose true score might be at the cut-point of 60% (39 points), i.e. passing. At this confidence level, just over a quarter of failing test takers (26.8%) are allowed to retake the test. Similarly, if we wanted to be 90% certain of not making a test taker with a potentially passing true score repeat the whole course, this would entail almost half the failing students (45.7%) resitting the exam, i.e. those with scores between 52% and 59%. Decisions on how wide such a safety margin should be invariably depend on the stakes of the test and the available resources. As already outlined above, the EBC2 exam is a medium stakes test, which means that the worst-case scenario for failing students is repeating the course, and they are entitled to do so up to twice. Under these circumstances, a 77% probability of not failing test takers outright whose true score might be a passing one seems to be sufficiently reliable and practicable, given the available resources.

	Raw score[1]	Proportion score	CI[2]	N	Fails CumN (Cum%)
Pass	39.0 - 65.0	60% - 100%	-	182	-
Fail, with resit	36.0 - 38.5	**55%** - 59%	$CI._{77}$	37	37 (26.8%)
Fail, no resit	34.0 - 35.0	**52%** - 54%	$CI._{90}$	26	63 (45.7%)
	32.5 - 33.0	**50%** - 51%	$CI._{95}$	17	80 (58.0%)
	0.0 - 32.0	0% - 49%	-	58	138 (100.0%)

Table 3: EBC2 scores and cut-off points

3.2. Criterion-related, concurrent validity

As outlined in the methodology section, we employed a differential groups design to investigate whether the EBC2 exam tests business knowledge as well as language skills. The multivariate analysis (MANOVA) discussed below aims to determine the impact of business knowledge on the test scores. For this purpose, the test takers were classified into two groups according to the extent of their business knowledge. This grouping variable serves as the independent variable in the subsequent multivariate analysis. The raw scores on the individual test sections represent the three dependent variables (see Tab. 4).

There was a statistically significant difference between students with high and low business knowledge on the combined dependent variables (combined test sections), $F (3, 314) = 8.608$, $p = 0.000$; Wilks' Lambda = 0.924; partial eta squared = 0.076 (see Tab. 4). When the results for the dependent variables (individual test sections) were considered separately, all differences reached statistical significance: Section 1, $F (1, 316) = 21.382$, $p = 0.000$, partial eta squared = 0.063; Section 2, $F = (1, 316) = 15.808$, $p = 0.000$, partial eta squared = 0.048; Section 3.1, $F (1, 316) = 9.960$, $p = 0.000$, partial eta squared = 0.031.

Variables	p	Partial eta squared
Independent variable		
Business knowledge		
Dependent variables (Section raw scores)		
Sections 1, 2, 3.1 combined	0.000	0.076 (7.6%)
Section 1 Business content and terminology	0.000	0.063 (6.3%)
Section 2 Language: grammar and vocabulary	0.000	0.048 (4.8%)
Section 3.1 Text comprehension	0.000	0.031 (3.1%)

Table 4: Effect of business knowledge on EBC2 scores (MANOVA - Wilk's Lambda)

This means that business knowledge had a significant impact on the overall test score, accounting for 7.6% of the variance (see Tab. 4), which represents a "medium effect size" (Cohen 1988: 22). At first glance, a confounding variable could arguably be at play here: After all, students with high background knowledge and high EBC2 scores might simply have superior study skills affecting both variables (business knowledge and EBC2 scores) independently. However, this

argument does not explain the differentiated effect of background knowledge on the individual test sections. Section 1 of the test covers *Business content and terminology*, and in this part business knowledge indeed shows the highest effect size, explaining 6.3% of the variance. On the other hand, effect sizes are smaller in Sections 2 (4.8%) and Section 3.1 (3.1%), which can be explained by the fact that these tap constructs less closely connected with business knowledge, viz. *Language: grammar and vocabulary* (Section 2) and *Text comprehension* (Section 3.1). These observations are clearly consistent with assuming that the EBC2 exam tests business knowledge as well as language skills and thus underscores its concurrent validity. Platzer and Zeilinger (forthcoming) report similar findings on the basis of a correlational approach.

4. Conclusion

Two key aspects of validity were investigated in the present paper, viz. scoring validity and criterion-related, concurrent validity. In terms of scoring validity, only item facilities were largely satisfactory, whereas the proliferation of discrimination values below 0.30 means that the majority of items would normally require revision. However, as such item revision remains unfeasible under the prevailing conditions, the resulting reliability estimate of 0.79 was respectable, but - in the final analysis - only marginally acceptable. It consequently seems clear that much would be gained in terms of scoring validity if proper piloting could take place, which - unfortunately - remains infeasible in the current setting. Nonetheless, we demonstrated that this shortcoming is compensated for by the fact that students who fail by a margin of no more than 5 percentage points are allowed to resit the exam, thus ensuring that we can be reasonably certain ($CI_{.77}$) that failing test takers whose true score might be a passing one are entitled to retake the test. We regard this approach to be both practicable and sufficiently reliable given that EBC2 is a medium stakes exam. As far as concurrent validity is concerned, a MANOVA confirmed that business knowledge has a significant effect on the EBC2 scores, with the biggest effect size in the section devoted to business content. In other words, the EBC2 exam taps business knowledge as well as language skills and thus tests the construct stipulated in the course aims. The EBC2 exam therefore addresses the relevant construct and is sufficiently reliable, despite the absence of pre-testing. However, this also implies that the rigorous test development process needs to be maintained in order to keep up this level of reliability.

5. Acknowledgements

We are grateful to Mr Schelenz of WU Academic Controlling for processing our test takers' grades on their business administration classes and for making the results available to us.

6. Notes

[1] Rounded to the nearest half mark.
[2] Confidence Interval for the **lower** bound of the cut-score.

7. References

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

Bachman, L. (2004) *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Brown, J., & Hudson, T. (2002) *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Carr, N. (2011) *Designing and analyzing language tests*. Oxford, New York: Oxford University Press.

Clapham, C. (1993) Is ESP testing justified? In D. Douglas & C. Chapelle (eds.), *A new decade of language testing research*. Alexandria, VA: TESOL Publications, 257-271.

Clapham, C. (1996) *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Cohen, J. (1988) *Statistical power analysis in the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Davies, A. (2001) The logic of testing Languages for Specific Purposes. *Language Testing, 18*, 133-147.

Douglas, D. (2000) *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Hughes, A. (2003) *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Ingham, K., & Thighe, D. (2006) Issues with developing a test in LSP. The International Certificate in Financial English. *University of Cambridge ESOL Examinations Research Notes, 25*, 5-9.

O'Sullivan, B. (2006) *Issues in testing Business English. The revision of the Cambridge Business English Certificates*. Cambridge: Cambridge University Press.

Platzer, H. & Zeilinger, R. (forthcoming) The Bermuda Triangle of ESP testing: test security - validity - field-specific content. In *Proceedings of The 6th International Language Conference on The Importance of Learning Professional Foreign Languages for Communication between Cultures, 19 - 20 Sept. 2013*. Celje: University of Maribor.

Robinson, P. (1989) An overview of English for specific purposes. In H. Coleman (ed.), *Working with language: a multidisciplinary consideration of language use in work contexts*. Berlin: Mouton de Gruyter, 395-427.

Thighe, D. (2007) Cambridge ESOL and tests of English for Specific Purposes. *University of Cambridge ESOL Examinations Research Notes, 27*, 2-4.

Weir, C. (2005) *Language testing and validation. An evidence-based approach*. Houndmills: Palgrave Macmillan.