Front. Educ. China 2014, 9(2): 238–249
DOI 10.3868/s110-003-014-0018-x

RESEARCH ARTICLE

David CONIAM

Pursuing the Qualities of a “Good” Test

Abstract This article examines the issue of the quality of teacher-produced tests, limiting itself in the current context to objective, multiple-choice tests. The article investigates a short, two-part 20-item English language test. After a brief overview of the key test qualities of reliability and validity, the article examines the two subtests in terms of test and item quality, using standard classical test statistics. Unsurprisingly, the pretested items outperform the teacher-produced test. The differences between the two subtests underscore issues about the quality (or lack thereof) of teacher-produced tests. The article ends with suggestions of how teacher-produced tests might be improved.

Keywords assessment, test quality, teacher-produced tests

Introduction

This article examines the vexed issue of English language tests produced by non-experts (in the main, non-native-speaking English language teachers) and the extent to which teacher-produced tests can attain acceptable quality.

Teacher-produced tests have long been a topic of concern among educators. Gronlund (1985, p. 267) discusses a number of salient differences between standardized and teacher-produced tests, stating that the latter may be more valid and relevant to the target test takers as long as they are well constructed. He notes, however, that the quality of test items on standardized tests is generally higher since they are written by specialists, pretested and refined, with reliability consequently very high. In contrast, the item quality of teacher-produced tests—which are typically put together quickly and used once only (Cunningham, 1998, p. 171)—is generally much lower, with

David CONIAM (✉)
Department of Curriculum and Instruction, The Hong Kong Institute of Education, Hong Kong, China
E-mail: coniam@ied.edu.hk

reliability much more uncertain.

Popham states that while a teacher's job is to teach—which many enjoy and do well—few like having to test (1995, p. 1), with tests constructed by teachers beset by eternal problems such as lack of time and resources. Regarding test quality, Kubiszyn and Borich (2003) state that many teacher-made tests "…are subject to various sources of error that impair their reliability of their scores and, consequently, their accuracy" (p. 303), in large part, due to inadequacies during teachers' tertiary education. Indeed, the fact that many teachers are not well versed with the principles of educational measurement, Popham rather appositely labels "assessment illiteracy" (2001, p. 26). Many educators have indeed noted that, for teachers, producing good tests is a demanding task, with good multiple-choice items, in particular, being extremely difficult to write (Hughes, 2003). It is, therefore, against this backdrop that the current study stands, investigating the quality of teacher-produced objective test items, in the context of tests (i.e., end-of-year) which Hong Kong teachers regularly have to produce.

Coniam (2009) reported a study with 31 Hong Kong EFL teachers enrolled on a three-unit course on Language Testing (forming part of a two-year part-time MA in ELT) where teachers-in-training—in a test design and analysis cycle lasting approximately six weeks (two to three times as long as teachers generally spend on test development considerations)—produced real tests aimed at their own students. Subsequent to administering the tests on their own students, they analyzed their tests with classical test statistics, and then commented on and refined their own tests on the basis of the item analyses produced from their students' responses. After such an extensive cycle, good reliability figures were nonetheless only achieved in half of the tests produced, with comparatively few of the items produced by participants deemed "good."

The current study extends the work of Coniam (2009) regarding the issue of quality in teacher-produced tests in a direct, simple, manner. Having obtained a ten-item test from a private school preparing Hong Kong students for the public examination, a 20-item test was then constructed. The first ten items were those from the private school test; the second ten were items from a calibrated database of the author's. What will be presented in the current article is therefore a contrastive picture of ten items produced by language assessment non-experts, and ten items produced and refined using a "quality-control" system involving pretesting, analysis and refining.

The Current Study: Scope and Background

This section describes participants, background to the Hong Kong school system and the methodology employed.

Background to the Hong Kong School System

Hong Kong has a 3 + 3 secondary school system with a single public examination (the Hong Kong Diploma in Secondary Education, or HKDSE) at the end of Year 12 (age 18), with an annual candidature of 80,000. Secondary schools are streamed into thirds in terms of ability, with Band One the most able, and Band Three the least.

Test Quality Issues

Since the HKDSE is an achievement test, the qualities of a "good" achievement test will now be briefly discussed. The public examinations body—the Hong Kong Examinations and Assessment Authority (HKEAA)—places great emphasis on the validity and reliability of the testing material used in its English language examinations. With regards to validity, the issues are how well the different parts of the test illustrate what a candidate can do in English, and how well test scores provide an indication of what candidates can do in English, along with test consequence and interpretations (Messick, 1989; Bachman & Palmer, 2010). In particular, the HKEAA investigates how well the test assesses what candidates have learnt, how well test content matches the target candidates—in terms of language, topic etc.—and how far the tasks have correspondingly relevant "communicative" contexts.

On the issue of reliability, any given administration of the HKDSE will be expected to return results similar to previous HKDSE examinations for similar cohorts of candidates (see Hughes, 2003, p. 36).

If a test is to be of high validity and reliability, it needs to be well constructed, with the features of a "good" achievement test involving the following principles (see Gronlund, 1985; Hughes, 2003):

- A test is at an "acceptable" level of difficulty (in the region of 0.5–0.6);
- Most candidates manage to finish the test;
- Candidates do their best on each item (i.e., they do not guess, nor draw on real-world knowledge or "extraneous" clues from other items or other parts of the test);
- The items discriminate, in that more able candidates get items right and less able candidates get them wrong.

Finally, items are "well-designed" (Falvey, Holbrook, & Coniam, 1994) in that:

- An item has only one key;
- All distractors do some work attracting (ideally less able) candidates;
- Errors do not contain grammatical, syntactic or spelling errors;
- All distractors are of similar length, with similar wordings used to maintain parallelism among the distractors;
- There are no semantic or grammatical clues between the stem and any of the distractors.

Test Quality Statistics

Methods of examining test quality in psychometric testing involve the use of classical test statistics such as the mean and standard deviation; internal test reliability using a statistic such as coefficient alpha; and item statistics such as the facility and discrimination indices (see Davidson, 2000, for a full explanation of "statistical hand tools"). These commonly-accepted statistics were those used in the Coniam (2009) study, and are drawn on again in the current project. The statistics will now be briefly described.

An ideal mean for a "final achievement" test (Hughes, 2003, p. 13) should be in the region of 0.5. Such a mean suggests that the test is generally appropriate to the level of a "typical" or "average" student in the class. A low mean can suggest that the test is too difficult, with a high mean suggesting that it is too easy (Burton et al., 1991). A mean in the region of 0.5 in general indicates that most students managed to finish it, i.e., that they did their best. Further, a mean of 0.5–0.6 indicates that student scores are spread out, and maximizes a test's discriminating power (Gronlund, 1985, p. 103).

Moving on to reliability, greater reliability is associated with test length, with the desirable level generally taken as 0.8 (for longer tests, i.e., with 80 or more items; see Ebel, 1965, p. 337). With shorter tests, lower reliability figures are cited; Ebel states 0.5 for 20 items and 0.33 for ten items (1965, p. 337).

While reliability is a useful starting point for determining a test's worth, the major tool accepted by many educators for determining objective test effectiveness is item analysis, which examines the contribution that each item makes to a test's worth. The usual classical statistics for classifying items as "good" or "bad" are the facility and discrimination indices. With regard to the facility index, an acceptable range is often taken as 0.3 to 0.8 (Falvey et al., 1994, p. 119). For the discrimination index, the key having a good value is taken as being 0.3 or better (Falvey et al., 1994, p. 126). A poor item is therefore one where the facility index is either too high (above 0.8) or too low (below 0.3), or more crucially where the discrimination index is lower than 0.3.

The Appendix 1 presents the ten items produced by the private school. Minor amendments—which do not compromise the integrity (or lack thereof)—have been made to each item to anonymize the source. The analysis of the items suggests there are structural problems associated with virtually every item.

Test and Subjects

The test in the current study was produced, as mentioned, by a private school preparing Hong Kong students for the Hong Kong HKDSE English language public examination, which students take at the end of Year 12. As preparation for this examination, many Hong Kong students attend private schools which give students examination tips and past examination paper practice, with much of the material produced in-house.

The ten-item private school test is an authentic test, which has been used completely "as is"; unamended and unaltered. The keys are those supplied by the school; these keys have been specified as answers in the item analysis conducted in the study. While a number of these keys are clearly faulty, they have been retained purposely to illustrate how poorly-designed test items impact on test validity and reliability.

It can be seen that the majority of the items are supposedly assessing test takers' lexical and collocational knowledge; items 9 and 10 are more syntactically-oriented in their focus. In the current study, the focus is not how well students fare at the testing points per se, but rather the teachers' test-design skills.

From the ten items presented in the Appendix 1, a 20-item test consisting of two ten-item subtests was constructed. The first comprised the ten items produced by the private school; the second ten items came from the author's "quality controlled" (i.e., pretested and refined) item bank. The second set of items matched those of the private school in terms of item type (multiple-choice), constructs tested (collocational-lexical, syntax) and ability level (Hong Kong Year 12). To an extent, then, the deck has been somewhat stacked against the test produced by the teachers in the private school by virtue of the fact that the counterbalance is a set of items whose quality has been verified, rather than inspired by teacher guesswork.

Statistically, a ten-item test (even 20) may be considered slightly suspect in terms of classical test statistics since the test length is so short. However, since the current study was attempting to directly research test quality, and like needed to be matched with like so in order to match the original ten-item test with one of similar makeup, the only option was to have two very short subtests of the same length.

The analysis below therefore shows a contrastive picture of ten items produced by a language assessment non-expert, and ten items produced and refined through a standard "quality-control" system involving pretesting and subsequent refining.

The composite test was run on two intact classes (N = 72) in a mid-ability (Band Two) secondary school. Given that the test was aimed at Year 12 students, for comparative purposes, the test was conducted with one mid-ability class of Year 11 students, and one mid-ability class of Year 12 students. The test duration was 15 minutes, and all students managed to complete all items in the test. The software used to conduct the analysis was Iteman v. 3.6. It should be noted that the student sample was not selected to compare Year 11 with Year 12 groups, but solely to provide data so that the hypotheses regarding test statistics could be investigated.

Hypotheses

The hypotheses in the current study are as follows:

1. Better overall test statistics (test mean, reliability [Cronbach's alpha]) will emerge from the set of "quality-controlled" test items than will emerge from the overall test statistics of the set of items produced by the private school;
2. Better individual item statistics (item facility and item discrimination) will emerge from the set of "quality-controlled" test items than will emerge from the set of items produced by the private school.

Results and Discussion

This section examines the results of the test from the perspective of the statistics outlined above. First a picture of the performance of the two participating classes is presented. An analysis of test-level statistics is then presented from the perspective of test mean and test reliability (Cronbach's alpha). This is followed by item-level statistics from the perspectives of item facility and item discrimination. A comparison of student performance in the two classes is presented below in Table 1 for reference.

Table 1 Class Means (max 20)

Class	*N*	*Mean*	*SD*	*t*-Test Results
Year 11	35	6.32	1.84	$t = -1.826$, df = 70,
Year 12	37	7.44	2.51	$p = .07$

As Table 1 indicates, the Year 12 class, as might be expected, scored higher than the Year 11 class, although at the 5% level, no significance emerged, and, as mentioned above, the comparative performance of the two groups per se was not being investigated.

Three sets of results will now be presented below. Each involves an analysis of the whole test followed by analyses of each subtest. Results are now presented for test-level statistics. As mentioned above, Ebel (1965) states that a desirable level of alpha with a 20-item test is 0.5, with a level of 0.33 for a ten-item test.

As can be seen from Table 2, the mean of the private school items was low at 0.2. In contrast, the "quality-controlled" items exhibited a more acceptable mean of 0.43, although it is accepted that this was still slightly on the difficult side.

Table 2 Test-Level Statistics

Statistic	Private School Items	"Quality-Controlled" Items
Mean	2.0/10 (0.20)	4.3/10 (0.43)
SD	1.27	1.79
Alpha	0.10	0.45

A similar situation can be seen with test reliability. The alpha for the private school items was a very low 0.10, considerably lower than the desired figure of 0.33. The "quality-controlled" items returned an alpha of 0.45, which is actually quite acceptable for only ten items. It will be appreciated that the reliability of the private school subtest is clearly poor.

Discussion now moves to statistics for individual items. In terms of item facility values, three "verdicts" are presented. A level of 0.3–0.8 is classified as "acceptable," below 0.3 as "difficult" and above 0.8 as "easy."

For discrimination purposes, two verdicts are presented. Correlations above 0.3 are classified as "good," and correlations below 0.3 are labeled "poor." Table 3 presents a summary of item facility and discrimination values, with verdicts.

Table 3 Summary of Item Facility and Item Discrimination Values

Statistic	Definition	Private School Items	"Quality-Controlled" Items
Item facility (IF)	Acceptable (0.3 – 0.8)	2	5
	Difficult (< 0.3)	8	3
	Easy (> 0.8)	-	2
Item discrimination (ID)	Good (>+0.3)	7	10
	Poor (< 0.2)	3	-

As Table 3 illustrates, item level quality was much lower with the private school items. In terms of item facility values, only two items had acceptable facility values as compared with five in the "quality-controlled" set. The picture was slightly better with item discrimination values, where seven of the private school items had good discriminations, although this still compared poorly with all ten items having good discrimination in the "quality-controlled" item set.

While item facility and discrimination may be analysed separately, it is when they are brought together to give a composite picture of "good" items that a more revealing picture of item quality emerges. "Good" items are defined (e.g., Falvey et al., 1994) as having an item facility in the range of 0.3–0.8 and an item discrimination above +0.3. Table 4 presents a summary of the results.

Table 4 Number of Good Items

Private School Items	“Quality-Controlled” Items
2	10

As can be seen, only two “good” items emerged on the private school test, as compared with the “quality-controlled” items, where all ten items could be classified as “good.”

What needs extended discussion—for which there is not space in the current article—is why the private school test items should be so poor. Appendix 1 presents a brief analysis of each item. As can be seen, each item suffers from a structural error in design: some items have two keys (items 3 and 7, for example); some items are rendered invalid because of language problems such as grammatical or syntactic errors (items 2 and 5); some items have unintended links to certain distractors (item 6, “a … inexpensive,” must be ruled out). The fact that there are problematic issues with all items severely compromise the validity of the ten private school items, validity problems which are underscored by poor item facility and discrimination values. A comparative analysis of the “quality-controlled” items has not been provided. In the main, this is because the focus is on why items are poor, rather than why items are acceptable. In the case of good items, there is actually not a great deal to say, apart from a somewhat bland comment such as: “The item worked well in terms of facility and discrimination; the distractors all functioned well.”

Conclusion

The hypotheses in the current study were that more acceptable overall test statistics (test mean, reliability) would emerge from the “quality-controlled” subtest than from the subtest produced by the private school, and that higher individual item statistics would be returned by the “quality-controlled” test items than by the private school items.

The results very clearly indicate that both hypotheses are accepted. On all measures, the “quality-controlled” items returned higher figures than the items produced by the private school. With regard to overall test statistics, where 0.5 was the target indicating general fit to the target subjects, the private school items mean was 0.2, compared with the figure of 0.43 for the "quality-controlled" items. More crucially, the alpha figure for reliability (where the target was 0.33 for ten items) was a very low 0.10 for the private school items, and 0.45 for the "quality-controlled" items.

In terms of individual item statistics, only two of the private school items could be classified as "good," as compared with all ten "quality-controlled" items.

From the two short subtests presented in the current study, it is clear that it is very easy to set bad tests. Much of the test practice material that emerges (often unmoderated)—from teachers and especially from teachers in institutions such as private "cram" schools—is actually a waste of time, and unfair to students in terms of the results generated.

The corollary of this is, however, that it is very demanding to set good tests; that is, tests which, a) are valid; b) fit candidates' broad ability levels; and c) provide a score which can be interpreted as an indicator of the (language use) ability that we want to measure. The question which then naturally arises is: Is there a solution?

The issue is complex, and has been discussed previously from a number of angles.

Educators such as Popham (2001) and Cunningham (1998) see it as axiomatic that teachers understand some of the basic principles of educational measurement. With this background, they argue, teachers have a better chance of recognizing, and thereby addressing, some of the pitfalls in the tests they produce.

Another issue is, however, time and support. Participants in the Coniam (2009) study made the point that test development in their schools was generally a solitary task. Little thought was given to test specifications, and tests were essentially cut and pasted from existing sources, with little reference to or moderation from other teachers.

In an attempt to solve the issue and in order to sustain good test design and analysis in a school setting, it is recommended that more than one teacher should be involved in assessment for each subject. They should be given time and resources through reduced teaching duties to properly accomplish this. Furthermore, it might be possible for staff development days, or days worked in the holidays to be used for these purposes.

The final issue considers whether professional training can have a long-term effect on the quality of teachers' tests. The fact that tests are often used only once and then discarded is a significant problem. Popham, for example, suggests that teachers build up an item bank of “good” subtests that might, in the long run, help to ease their test production burden (2001, p. 108).

It is clear that a great deal of continuing support and professional development is still required in the area of assessment and test design for the majority of teachers of English in Hong Kong and elsewhere (c.f. Carter, 1984, who describes similar needs in four states in the USA) as an integral part of their professional self-development. Some of the suggestions above, if carried out purposefully, can mitigate the poor knock-on effects of teacher-made tests.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Retrieved February 18, 2014, from http://testing.byu.edu/info/handbooks/betteritems.pdf

Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education*, *35*(6), 57–60. doi: 10.1177/002248718403500613

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the impact of training in test development principles on improving test quality. *SYSTEM*, *37*(2), 226–242. doi: 10.1016/j.system.2008.11.008

Cunningham, G. K. (1998). *Assessment in the classroom: Constructing and interpreting texts*. London, England: Falmer Press.

Davidson, F. (2000). The language tester’s statistical toolbox. *System*, *28*(4), 605–617. doi: 10.1016/S0346-251X(00)00041-5

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students*. Hong Kong, China: Longman.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching* (5th ed.). New York, NY: Macmillan.

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

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice*. New York, NY: Harper Collins.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed.; pp. 13–103). Phoenix, AZ: American Council on Education/Macmillan.

Popham, W. J. (1995). *Classroom assessment*. Needham Heights, MA: Allyn & Bacon.

Popham, W. J. (2001). *The truth about testing: An educator’s call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

APPENDIX 1: Private School Items

1. His salary is ______ to support three persons, so he needs not bother himself with his living. a) enough b) sufficient c) adequate [**KEY**] d) plentiful **→A or B; but “needs” hence none**	6. He asked for a ______ price for the second-hand bicycle. a) cheap b) inexpensive c) low d) modest [**KEY**] **→A or C; B impossible**
2. He has made ______ investigation. a) adequate b) sufficient [**KEY**] c) ample d) plentiful **→Should be “investigations”, hence none**	7. She felt very ______ among these new colleagues she had never met. a) strange b) odd [**KEY**] c) queer d) peculiar **→A or B; D?**
3. The supervisor’s instructions are not very ______ . a) accurate b) correct c) exact d) precise [**KEY**] **→all, for different reasons**	8. The ______ of sleepless nights made her feel rather ill. a) strain [KEY] b) stress c) tension d) pressure **→A, B, D for different reasons?**
4. All scientists have to be ______ in making instrumental tests. a) accurate b) correct c) exact d) precise [**KEY**] **→“making instrumental tests”?; hence none**	9. “Where is my scarf?” “It’s on the ______ .” a) kitchen counter [**KEY**] b) kitchen’s counter c) counter kitchen d) counter’s kitchen **→A or B**
5. Small cars are ______ as compared with cars of standard size. a) cheap b) inexpensive c) low-priced [KEY] d) modest **→“as compared with,” “of standard size” poorly phrased; if we get past this, A or B; possibly C**	10. Can you please tell me the ______ for the Physics Lab and Psychology lab? a) room number b) room numbers [**KEY**] c) room’s number d) room’s numbers e) rooms’ number f) rooms’ numbers **→A or B (if 2 labs in one room)**

Notes: * Items slightly amended by the author; ** The keys are the school’s, the commentary the author’s.