A discussion on the threat posed to oral test validity by participant patterns

Simon EVANS

Put simply, a good test should, irrespective of discipline, define what ability it wishes to test, and then ensure that this trait is actually assessed. This describes the notion of validity, which is the gathering of *a priori,* or pre-test evidence in the shape of theoretical and content justification, and *a posterior,* or post-test in the form of statistical analysis of test scores. All good tests irrespective of size or gravity of decisions made should exhibit validity, though in most teaching contexts a trade-off between validity and practicality is the norm.

In the field of EFL test development, the widely held belief that validity of a test, or more specifically, validity of the inferences test-takers (TTs), once tested and graded, are subjected to, form the backbone of that test's value still pervades (e.g. Fulcher & Davidson, 2007). Indeed, this concept continues to represent the core of language testing research (Xioming, 2008; 177), the focus of which falls on elements that threaten a test's worth, which reinforces Hughes (1989) earlier declaration of the importance of validity when searching for solutions to language testing problems (p.22). Through in the case of oral testing, interaction should be a key feature (Weir: 1990; 73), interlocutors (TTs and examiners/assessors) are possible sources of *a posterior* problems (test bias).

In Weir's (2005) model of operational validity for all four language skills, the role of interlocutors in oral assessment is given saliency through its inclusion as a subsection of factors that influences context validity (p.XX); the "how" of testing. Therefore, it is self-

evident that the nature of the interaction between test protagonists can ultimately have a detrimental effect on educators' judgements of learners. Backman (1990) distinguishes between so-called systematic and random 'sources of error', which he classifies as test-taker characteristics (TTCs) (p. 164). Systematic errors, which are assumed to affect performance regularly (ibid) are listed in Fig.1. O'Sullivan (2006) reminds us that it is research which has found, and continues to find a multitude of TTCs that may place the validity of test score inferences in jeopardy (pp. 26-27).

Fig.1. Test-taker characteristics

Age
Sex
Cognitive style
Personality
Education level and general knowledge
Motivation to take test
Native language
Test awareness and preparation
Ethnic background

Adapted from O'Sullivan (2006)

By limiting acknowledgement of their effects on performance of tasks to those which demand limited or no interaction (ibid: 27), he is alluding to the possible sources of error, or test bias, that are limited to within the permutations of interlocutor discourse; examiner – TT, and TT – TT.

Issues with Oral test participant permutations

Before beginning our discussion of the advantages and disadvantages of each permutation in terms of participants' effects upon each other, I shall define the two combinations mentioned above;

1. **Examiner and test-taker.** In this context, the examiner will

attempt to exercise control over learner output by standardizing input, commonly in the form of questions.

2. **Test-taker and test-taker.** Here, there will be learner interaction with minimal input from the examiner.

A clear advantage of (1) its ability to ensure what is elicited covers the taught syllabus. With an examiner-controlled interview, an interviewer frame can be employed and there is greater chance of question uniformity, hence enabling greater scoring validity However, this is likely to come at the cost of authenticity, for this format is unlikely to be encountered in real life (Weir: 1990;76). Egyud & Glover (2001) add further caution by stating, "It would be wrong to choose exam formats that reflect the unrealistic interaction patterns common in teacher-centred classrooms" (p.75). An alternative is to limit the teacher to a management/observation role, i.e. impart test instructions and then grade TTs output. In oral testing, Hughes (1989) insists TTs be put at ease by the examiner in order to facilitate optimum performance (p.106). However, a study of 36 female Japanese junior college TTs concluded that the very presence of a native-speaker examiner overrides all other TT characteristics (Berry; 2007:143) .

There exists plenty of support for (2). O'Sullivan (2002) mentions the existence of widespread anecdotal evidence of teachers' support of the idea of better TT performance in pairs (p.279). Bennett (2012) declares that, "Pair testing is apparently well accepted as a valid method of assessing oral ability in learners of other languages" (p. 337). In an albeit limited study of secondary school EFL learners in Hungary, Egyud & Glover (2001) found evidence of learner's liking of the paired format ,that it helps students to produce their best, and are preferable to the examiner – TT format (p.70). Brown (2004) says the advantages of TT pair interviews as they allow for more TTs to be assessed in any given period of time (p.171), ideal for large classes, plus,

the interaction inherent in the test results in greater authenticity, which McGinley (2006) characterizes as a more natural flow to conversation (p.276), and Saville & Hargreaves (1999) as more possible speech patterns (p.44) . If the nature of elicited speech is the strength of the pair format then inter-TT relationships is its achilles heel. Hughes (1989) states that in the case of interaction with peers, one TT's performance is likely to be affected by the other's (p.105). To add credence to this assertion, in a limited study of Japanese learners, Norton (2005) found evidence of TT scaffolding of linguistic output when one candidate used the item 'otherwise' immediately after their partner had produced it (p.291). A drawback to the paired format is the danger of imbalance in conversation contributions performance (Weir: 1993; 35; Brown: 2004; 171), meaning a possible shortage of elicited language by which to assess ability (Foot: 1999; 37) . Norton (2005) concludes that TT performances in pairs is effected by the issue of acquaintanceship (p.287) and the amount of assessable output. As a result of a limited study, Ikeda (1998) in Saville & Hargreaves (1999) believes that TT anxiety can be reduced by permitting TTs to choose their own test partner. O'Sullivan (2002) mentions another Japanese study that produces evidence which suggests that TT scores were higher when candidates were placed with a friend (p.290). In a critique of paired oral tests, Foot (1999) challenges the intuitive assertion that knowing your partner relaxes the candidate by wondering whether in fact a friend in a test would attempt to negotiate meaning, for this could serve to illustrate their partner's weaknesses (p.37). In pairs, TTs spend too little time talking (Foot: 1999; 40) , therefore calling into question the paired format's effectiveness with lower level learners (ibid: 41) Egyud & Glover dispute (2001: 72). Teachers are sceptical about paired tests because of differing abilities within the pair (Bennett: 2012; 337).

In summary, it would appear that examiner – TT and paired TT formats have both positive and negative effects upon a test's validity. The former, while enabling output which appears in the syllabus may

satisfy the content aspect of Weir's (2005) context validity, does so by sacrificing naturally occurring discourse. The latter, with its more natural conversation pattern does likewise for the interlocutor element but then introduces the possibility of imbalance in linguistic output due to issues of acquaintance or differing abilities, thus calling into question the paired format's ability to score TT's fairly. This leads me to conclude that it is a tests purpose, or more specifically, which element (s) of communicative competence do we wish to test that should be the determining factor in choosing oral test format. If ability to produce language covered in the syllabus is the sole goal, often the case when assessing lower level students, then the examiner – TT format holds the advantage, while if the ability to produce speech for multiple functions has primacy, the paired format is better suited.

Word count: 1,298

References

Backman, L. F. (1990). Fundamental Considerations in Language Testing. OUP, Oxford.

Bennett, R. (2012). Is Linguistic Ability Variation in Paired Oral Language Testing Problematic? *ELT Journal Volume 66/3*.

Brown, H. D. (2004). Language Assessment: principles and classroom practices. Longman. New York.

Foot, M. C. (1999). Reply to Saville and Hargreaves. *ELT Journal Volume 53/1*.

Fulcher, G. & Davidson, F. (2007). Language Testing and Assessment. Routledge. New York

Hughes, A. (1989). Testing for Language Teachers. Cambridge University Press. Cambridge.

McGinley, K. (2006). The Test of Interactive English. *ELT Journal Volume 60/4*.

Norton, J. (2005). The Paired Format in the Cambridge Speaking Test. *ELT Journal Volume 59/4*.

O'Sullivan, B. (2002). Learner Acquaintanceship and Oral Proficiency Test Pair-Task Performance. *Language Testing Volume 19*

O'Sullivan, B. (2006). Modelling Performance in Tests of Spoken Language. Frankfurt. Peter Lang.

Saville, N & Hargreaves, P. (1999). Assessing Speaking in the Revised FCE. *ELT Journal Volume 53/1.*

Weir, C. (2005). Language Testing and Validation: an evidence-based approach. Palgrave. Oxford.

X, Xi. (2008). Methods of Test Validation in *Encyclopedia of Language and Education. eds. Shohamy, E. & Hornberger,* N. H. Springer. New York.