Vol. 1 (2009) 46-54 University of Reading ISSN 2040-3461
LANGUAGE STUDIES WORKING PAPERS
Editors: L.J. O'Brien and D.S. Giannoni

Reliability as a Context-Dependent Requirement for Writing Proficiency Assessment

Daniel Muñoz

In the following paper I provide an argument in favour of reconceptualising the role of reliability as a necessary requirement of writing assessment used as an observational procedure. The argument elaborates on the observation that the traditional approach to the role of statistical reliability estimation in writing assessment, which considers it a necessary requirement, is actually in conflict with real language/writing assessment contexts in which this estimation is either unfeasible or less relevant. Reasons for this situation may be found in the characteristics of reliability estimation and its interpretation in relation to the nature of writing and the characteristics of actual assessment procedures. I will correspondingly suggest that the traditional dilemma between validity and reliability does not hold for local contexts where in practice reliability measures may often be non-existent. It will be argued instead that an alternative concept of reliability as a context-dependent property (as opposed to a necessary condition) of writing assessment could account more adequately for the validity of assessment procedures in local contexts where reliability estimates may not be feasible or relevant.

1. Introduction: the nature of reliability

Reliability is a construct imported into language assessment from fields such as psychometrics and educational research (see, for example, Traub & Rowley 1980; Henning et al. 1985; Fan 1998; Harvey & Hammer 1999; Johnson et al. 2001; Rudner & Schafer 2001; Wiberg 2004; Steyer [online]). Interest in the conceptualisation of reliability in the field of writing assessment dates back to early observations by Edgeworth (1890) regarding inconsistency in the ratings of different examiners for a particular writing task. This attention has been recently enhanced by increasing interest in the development of writing abilities as a key factor in the academic and professional success of individuals (Weigle 2002; Hamp-Lyons 2003).

The rationale behind the concept of reliability derives from the conceptualization of variation and its relation to the observation process: any measurement constitutes an act of observation which depends on several factors that will vary for several reasons and to different extents (Slomp & Fuite 2004). If critical factors such as the observer(s), the occasion(s) of observation, the object(s) of observation and the actual method of observation change from one instance to another, differences in the final observation will result. Human observations are thus naturally error-bound and context-specific (Skehan 1984; Davies 1988; Fan 1998).

In language assessment, this natural variability will be also determined by other particular variables: the native speaker (i.e. the native model targeted by a test), cut-off criteria, scoring criteria, test design and the language description model supported by test designers, all concepts which can be considered "ill defined and subject to unreliability" (Davies 1988: 32). Since language assessment is always subject to some degree of variation (i.e. unreliability), ensuring as much reliability as possible becomes a natural concern in specific areas such as writing assessment. This concern is simultaneously theoretical, technical and ethical, to the extent that a complex system of decisions regarding an individual's future will be partly based on the use and interpretation of her/his writing skills (Weigle 2002; Hamp-Lyons 2003; Casanave 2004).

Unfortunately, as I intend to show in this paper, the definition of reliability still requires substantial refinement as to the role it has in the assessment of both language and writing, particularly regarding its relation with the construct of validity. The need for conceptual refinement may be also extended to the actual estimation of reliability, its interpretation and the criteria for its acceptability.

Following the controversy regarding the validity/reliability relation (Davies 1988; Casanave 2004; Slomp & Fuite 2004), I will address general conceptual and technical difficulties in the application of the reliability construct in writing assessment, most of which arise from the unsystematic nature of observational procedures and of writing assessment itself. I will also argue that, in view of these difficulties, reliability should not be considered as an inherent requisite for *all* writing assessment procedures but rather as a context-dependent property. The core argument will be that reliability is a key property for global writing assessment contexts in which little or no individual information about writers is available and/or where there is a strong need for public accountability. On the other hand, reliability will be an optional requisite in local writing assessment contexts, often characterised by individual information on writers and less accountability pressures.

The ensuing discussion is related to the primary definition of reliability and its general conceptual implications, so finer distinctions between, for example, L1/L2 writing assessment/test/measurement will not be made. Finally, although the discussion is focused on writing assessment, I am assuming that the main points of my argument hold true also for the assessment of other language skills, as long as they are equally based on data drawn from human observations. I believe that improvement regarding these conceptual controversies is relevant inasmuch as it could contribute to an increase in the validity and fairness of the writing assessment procedures based on these conceptual grounds.

2. Statistical reliability as a core property of assessment procedures

The issue of reliability and its special relation to validity constitutes the focus of an established sub-field of research in writing assessment (Hamp-Lyons 2007). According to Weigle (2002: 49), validity has been traditionally defined as *construct validity*, i.e. a test's capability of "measuring what it is intended to measure". As a result of the extensive work by Samuel Messick, the concept of construct validity has become more complex, now including also ethical considerations as to the consequences of test applications (Davies 1988; Hamp-Lyons 2003). Today a valid test should satisfy requirements regarding "the quality of a test instrument, its appropriacy for its intended purpose, and the potential for misuse of the test" (Hamp-Lyons 2003: 167). Further links with the concept of validity will be discussed later. It will be noted, for the time being, that the construct is characterised as a necessary (but not sufficient) condition for *any* kind of assessment.

Reliability, on the other hand, can be understood as "the ability of the test scores to be replicable – for example from one test occasion to another, or from one easy prompt to another" (Hamp-Lyons 2003: 163). Such a definition implies four key implications:

- Reliability is a statistical property of test scores
- It implies consistency in the scores at different occasions of assessment
- It implies some way to measure that consistency
- It constitutes, like validity, a necessary (but not sufficient) condition for *any* kind of assessment

Since any observational process is subject to differences that cannot be explained in relation to what was being measured (Edgeworth 1890; Henning et al. 1985; Rudner & Schafer 2001; Slomp & Fuite 2004; Bodoff 2008;), the extent to which the variations observed can be accounted for as having a systematic cause will then determine the level of reliability (or accuracy) of a procedure (Oller 1979; Traub & Rowley 1980; Krzanowski & Woods 1984; Rudner & Schafer 2001; Slomp & Fuite 2004).

The statistical estimation of reliability is thus aimed at determining "how much of the variance in a test on one occasion can be expected to be present on another occasion. In other words, reliability is essentially a problem of how well a given test can be expected to correlate with itself" (Oller 1979: 64). The term *variance* is here used to characterise the differences that scores may show in relation to one particular test. According to Krzanowski and Woods (1984: 1), reliability would then be "the extent to which differences in ability are *not* obscured by random measurement errors" and would be concerned with "how well a test measures *whatever it may be measuring*."

Language assessment has traditionally supported the centrality of both validity and reliability as definitional requirements of assessment procedures (Oller 1979; Traub & Rowley 1980; Krzanowski & Woods 1984; Davies et al. 1999; Fan 1998; Underwood & Murphy 1998; Wiberg 2004; Bodoff 2008). Indeed, the first article published in the *Language Testing Journal* discussed precisely the issue of the statistical treatment of reliability (Krzanowski & Woods 1984). This general view is also illustrated by the *Dictionary of Language Testing*'s definition of reliability as "the actual level of agreement between the results of one test with itself or with another test. Such agreement ideally would be the same if there were no measurement error" (Davies et al. 1999: 168).

The centrality of reliability as a defining characteristic of writing assessment in particular also offers grounds for disagreement. Casanave (2004) characterises this disagreement as a controversy between an 'objective' and a 'subjective' view of assessment. From an 'objective' perspective, reliability is considered an integral part of validity: if consistency across and along ratings cannot be guaranteed, there is no possibility of ensuring that what is purported to be measured is being actually measured. Fair, 'objective' assessment implies measures and judgements uncontaminated by personal bias from test designers and raters.

This rationale, explains Casanave, has supported the extensive research on reliability of the last decades. An extreme version can be identified behind the use of indirect writing-assessment, now discarded in favour of direct testing on account of its greater construct and face validity (i.e. that they actually measure writing proficiency so that non-specialists would consider such measures as valid) (Casanave 2004; Hamp-Lyons 2003; Weigle 2002)[1].

In summary, reliability in language/writing assessment is generally characterised by being both a statistically-based construct and a *sine qua non* requisite for assessment. As I will discuss later, these definitional criteria are often incompatible with the actual nature of writing assessment in real contexts.

3. How reliable is reliability in writing assessment?

According to the American Statistical Association, statistical analysis can provide "crucial guidance in determining what information is reliable and which predictions can be trusted" (ASA 2008), which implies that a reasonable expectation for tests is that we can 'trust' their results, as they will inform our decisions.

However, the estimation of reliability in actual writing assessment procedures is not as straightforward a process as this definition may lead one to believe. This is mainly due to the variable nature of writing and of the assessment procedures themselves (Oller 1979; Traub & Rowley 1980; Krzanowski & Woods 1984; Underwood & Murphy 1998; Slomp & Fuite

[1] It is interesting to notice that this movement did not necessarily mean a reaction against the concept of reliability but rather towards ensuring the validity of assessment procedures.

2004). Indeed, it is *because* of this variability that the role of statistical analysis becomes that of determining "how much of the variance in any given test can be attributed to the construct the test is supposed to measure" (Oller 1979: 68), i.e. of determining validity. Two relevant sources of random error (or variation) can be identified in writing assessment: human unsystematicity and the actual estimation of reliability.

3.1. Human unsystematicity

Identifying systematic differences in an individual's writing is in itself a problematic task. Writers perform differently on different occasions (Skehan 1984; Weigle 2002; Hamp-Lyons 2003; Brown et al. 2004; Casanave 2004) and the possible contextual causes (room temperature, noise, available infrastructure, etc.) or personal causes (tiredness, anxiety, personal background, etc.) for variation are difficult to account for.

Simultaneously, writing teachers' and raters' decisions will inevitably be shaped to some extent by their personal beliefs, preferences and biases (Rudner 1992; Underwood & Murphy 1998; Weigle 1998, 2002; Johnson et al. 2001; Brooks 2004; Elder et al. 2007; Eckes 2008). Differences in judgement will respond to testing conditions and individual features which, again, are difficult to evaluate.

3.2. Difficulties in the estimation of reliability and its application

Writing is technically a very complex observational object: several levels of language and discourse analysis must be taken into account (Davies 1988; Stansfield & Ross 1988; Hamp-Lyons 2003, 2007) and observational categories will vary substantially according to the selected descriptive framework for language (i.e. grammar) (Skehan 1984; Davies 1988; Hamp-Lyons 2003; Casanave 2004).

Additionally, procedures and formulas to estimate reliability are also heterogeneous. These include, for example, the test-retest method, split-half method, alternate forms technique, Cronbach's Alpha, Kuder Richardson 20 and 21, and the Spearman Brown formula (Oller 1979; Traub & Rowley 1980; Krzanowski & Woods 1984; Rudner & Schafer 2001; Bodoff 2008), all of which work roughly on the principle of estimating the correlation between different instances of assessment (Oller 1979). However, as Krzanowski and Woods (1984) warn, each of them apply to different assessment conditions and then do not necessarily lead to comparable interpretations.[2]

The acceptability level of reliability in a writing assessment is another area of concern. The popular 0.8 coefficient, for example, is indicated as a safe relative estimation of inter-rater reliability for most educational contexts (Traub & Rowley 1980; Davies 1988; Johnson et al. 2000; Penny et al. 2000; Hamp-Lyons 2003). However, assuming by default that there is indeed need for a reliability coefficient overlooks the fact that high inter-rater reliability may not necessarily ensure higher validity. Indeed, Hamp-Lyons (2007) provides evidence from rating procedures exhibiting qualitatively different judgements and criteria among otherwise highly reliable groups of writing testers.

The actual application of these estimations poses an additional concern. First, practitioners are seldom aware of the subtle but ultimately relevant complexities of reliability estimations (Krzanowski & Woods 1984) and the same can be said of other stakeholders, especially those less directly involved with the classroom context (e.g. administrators and policy-makers)[3].

[2] It is also interesting to note that a variance test analysis (e.g. ANOVA), is not necessarily the preferred option in the field (Oller 1979; Traub & Rowley 1980; Krzanowski & Woods 1984; Rudner & Schafer 2001; Bodoff 2008), probably due to the technical complexity of this procedure (Krzanowski & Woods 1984).

[3] The implications of having decision-making processes systems informed by writing assessment results that are not properly interpreted should also be a matter for concern (Hamp-Lyons 2007).

Additionally, in many local classroom contexts, a focus on statistical reliability measures may be impractical due to time, money and other practical constraints. Inter-rater and intra-rater reliability can be discarded, for example, when there is only one teacher with little time to act as a rater. The adoption of a statistically-based construct of reliability thus implies inherent problems when applied to language/writing assessment practices, namely:

- Direct writing assessment is inherently unreliable because it is an observational process and because of the nature and circumstances of writing.
- No shared framework exists yet to either measure reliability or to judge the acceptability of those measures for writing assessment.
- No shared language/discourse categories exist on which educational stakeholders can interpret assessment judgments.
- There are many (especially local) contexts in which reliability estimation exceeds available resources.
- Relevant stakeholders may not be aware of the uncertainties implied by different reliability estimation procedures.

As I will show later on, this situation has important implications as to the central status of reliability as a necessary condition for *any* writing assessment context.

4. Reliability and validity: the 'objective' and the 'subjective' views

As explained earlier, Casanave (2004) observes a general difference between an 'objective' and a 'subjective' approach to reliability in writing assessment. Table 1 summarises some of the more specific and relevant differences between the two.

Objective approach to reliability	*Subjective approach to reliability (Weak version)*	*Subjective approach to reliability (Strong version)*
Reliability is a necessary condition for test validity	Reliability is an optional requisite of validity	Reliability undermines validity
Underlying large-scale tests	Underlying double marking	Underlying portfolio assessment
Raters' individual judgement decreases reliability levels	Raters' individual judgement should not affect reliability levels	The effect of raters' individual judgment on reliability is a secondary issue
Feasible in global contexts	Feasible in local contexts	Feasible in local contexts
Unfeasible/unattractive in local contexts	Unfeasible/unattractive in global contexts	Unfeasible/unattractive/irrelevant in global contexts

Table 1. Subjective versus objective approaches to writing assessment reliability (based on Casanave 2004).

The objective approach assumes that reliability is a condition to be fulfilled by writing tests so that they can be valid, in the sense that inconsistent results cannot possibly inform about an evaluation target. This view is held by most writing assessment researchers (see Oller 1979; Krzanowski & Woods 1984; Hamp-Lyons 2003). It is also the rationale that underlies current practices of large-scale, high-stakes writing tests such as IELTS or TOEFL (see Weigle 2002) and thus their focus on ensuring score reliability and refining reliability estimations (Traub & Rowley 1980; Stansfield & Ross 1988; Brown et al. 2004; Casanave 2004).

An example of this focus is the effort to characterise and improve inter-rater reliability by means of several forms of standardisation training (see Weigle 2002). The aim of such procedures is to reduce the effect of individual, unsystematic variation in order to have similar ratings among different raters. This is normally achieved by providing detailed 'objective' descriptions of targeted outcomes (*rubrics, scales*) which raters are trained to follow (see, for example, Shrout & Fleiss 1979; Rudner 1992; Underwood & Murphy 1998; Weigle 1998; Stemler 2004; Elder et al. 2007; Eckes 2008). Finally, the objective approach is sensitive to the practical uses of tests and the pressure to obtain results for which examiners can be accountable to other stakeholders (see Weigle 2002; Hamp-Lyons 2007). High levels of statistically-validated score reliability are thus an understandable requirement for decision-making processes by educational policy-makers and institutional administrators.

Still following Casanave (2004), the subjective approach to writing assessment, views reliability and validity as two *independent* properties of assessment procedures and suggests that a valid assessment procedure should welcome raters' individual (statistically unreliable) responses to examinees' writings (Weigle 2002; Brooks 2004; Casanave 2004). Two positions may be held within this view. One is that reliability is necessary and can be ensured without necessarily suppressing raters' subjective responses, as in the case of double-marking. The other is that reliability among raters may be treated as a less relevant condition as compared to the validity of the assessment procedure (McKay 2007), as in portfolio assessment (Hamp-Lyons 2003).

The standard double-marking procedure relies on the holistic assessment of two raters who score the tests after one or two readings. The high levels of inter-rater reliability that may be achieved with this procedure, however, meet two drawbacks: the lack of detail as to specific features of test samples and the unattractive cost-benefit ratio of the procedure for large-scale programmes (Brooks 2004; Casanave 2004).

In portfolio assessment, on the other hand, no serious attempt to make judges agree is required, which seems to respond adequately to one-rater classroom contexts. Most importantly, portfolio assessment is also informed not only by demographic data but also by more direct qualitative information about students/writers (e.g. educational background, personal preferences, learning habits, motivation), which is not usually available in large-scale writing assessment (Weigle 2002; Hamp-Lyons 2003; Casanave 2004).

This subjective approach has two relevant implications for a conceptual discussion of the definitional role of reliability in writing assessment. In its weakest form, this implies that some writing assessment procedures may not need to be reliable to be valid. In its strongest form, it implies that statistical reliability may even be an indication of invalidity. Indeed, to the extent that subjective judgements are suppressed (as in standardisation training), the judgements of raters would inevitably ignore relevant sources of potentially useful qualitative, more personal information about writers.

However, the validity/reliability controversy may only be apparent, as Casanave (2004) explains, since each approach is suitable for a different context. Subjective assessment may be more useful for local contexts, such as classrooms, and objective assessment for large-scale assessment. In fact, it may be argued that large-scale tests need to be highly reliable *precisely because* they have no access to other sources of information to enhance their validity. This means that the dependence of validity on reliability may actually be a function of the context rather than a necessary definitional condition. In practical and conceptual terms, this dependence seems more appropriate for large-scale testing rather than classroom contexts, on account of differences in the type and quality of information available for assessment.

5. Discussion

The picture so far depicted is that the definition of reliability as a necessary property of writing assessment and the characteristics of its estimation and application need further discussion. The argument so far can be summarised as follows:

- Reliability is a statistical measure of the extent to which we can trust the results of a given writing test. Reliability estimations and their interpretations will vary according to particular assessment contexts and purposes.
- In real practice, high score reliability is indeed necessary and feasible in large-scale assessment, where the actual tests are the main source of interpretable information. In

more local contexts (e.g. classrooms), however, reliability measurements are to a great extent unfeasible and, simultaneously, other sources for qualitative evaluation are available.

Therefore reliability measurements from large-scale writing assessment seem less useful for classroom contexts and vice versa. For the former, reliability is achieved by reducing raters' subjective judgement through the use of preset 'objective' descriptions to guide their judgments (i.e. scales). Agreement regarding these descriptions is thus interpreted as a sign of validity. For the latter, the validity of an evaluation is importantly based on the teacher's own direct, and often less formal, observations. The need for high reliability as an indicator of validity can then be seen more as a weakness of large-scale procedures inasmuch as relevant direct observations are unavailable.

In cases where most of the variance in language testing happens to be random or reliability estimates do not satisfy desirable levels, a heavy reliance on these estimates might simply be misleading as indicators of particular assessment observations. This leads to the consideration of the popular validity/reliability dilemma, which is normally presented as an option between either increasing levels of validity at the expense of reliability or vice versa (cf. Davies 1988; Hamp-Lyons 2007). Slomp and Fuite (2004: 197) consider a trade-off relationship as inevitable, due to the operation of an *uncertainty principle* according to which every observational process is subject to a loss of reliability which is inversely proportional to the loss of the perception of the observed object as a whole. As when using a microscope, more detail means more accuracy (reliability) but less perception of the object as a whole (validity); conversely, a broader view of the object implies less access to detail.

The validity/reliability balance will thus depend on the 'size' of the construct. A 'small' construct (e.g. the correct use of capital letters) will imply small differences between validity and reliability adjustments; a 'bigger' construct (e.g. academic writing proficiency) will determine greater loss of both validity and reliability. Slomp and Fuite (2004: 204) suggest then that, instead of focusing on whether to enhance either validity or reliability, we should aim at balancing them in order to improve the general *quality* of a writing test. As an integrative concept, this proposal agrees with Hamp-Lyons' (2003) idea of *quality*. Her solution to the dilemma, however, is to integrate reliability as part of the 'construct validity' concept so that, again, more reliability will enhance validity.

The impression, however, is that the validity/reliability dilemma presented so far fails to capture the conflict between the definition of reliability and actual writing assessment practices. As explained, although reliability is normally defined as a central requirement for *any* writing assessment procedure, it is in practice an optional, less relevant and often inexistent characteristic in many local contexts (for example, in the case of classroom double marking and portfolio assessment). In such contexts, no balance or integration is possible since one of the elements, i.e. reliability, would simply be missing. As an alternative viewpoint, I suggest that we should assume that the validity/reliability dilemma – as presented by Slomp and Fuite (2004) – is real, so that increasing levels of reliability will inevitably result in the loss of validity and vice versa.

Secondly, like Casanave (2002) we should consider reliability and validity as two separate properties of a test, of which validity is the *sine qua non* property of writing assessment and reliability is an optional, *context-dependent* property. In this view, the focus is still on validity as the main requisite of writing assessment, so that score reliability becomes a necessary requisite for assessment contexts without access to qualitative information from examinees and/or contexts that involve broader audiences to which examiners are accountable.

The loss or absence of reliability estimations in local contexts, in contrast, may not be an issue insofar as validity is highly increased by the teacher's more personal knowledge of students' development and also as the use of reliability estimations is often unfeasible. The approach also allows to better explain those contexts where reliability measurements may not be possible but where writing assessment *could still be considered valid* in virtue of the access to relevant qualitative information. In such cases, teachers and educational administrators may be relieved of the pressure of accommodating their teaching to costly procedures devised for different assessment conditions.

6. Conclusion

In this paper I have attempted to show that the definition of writing assessment reliability requires refinement in both its conceptual and technical aspects. The main observation is that, although consistency in the direct assessment of writing is important, we still need agreement as to what this consistency is, how it can be estimated, how much of it is desirable, and when it is most necessary. Another important point is that issues of statistical reliability are of key importance for the 'objective' assessment of writing (e.g. large-scale tests), whereas they may not be so relevant in more local contexts (e.g. classroom assessment), because estimation of reliability is either unfeasible and/or simply not required as a validity indicator.

I have then argued that there is a conceptual conflict between assuming that writing assessment must be reliable to be valid and, simultaneously, observing that statistical reliability estimations may not be needed to ensure writing assessment validity in particular (especially local) contexts. This conflict leads to a situation of incompatibility between the interpretation of local writing assessment and global writing assessment, to the extent that their validity is interpreted on different bases. A solution to this contradiction is a position in which no validity/reliability balance or integration is required. I suggest instead that both concepts should be kept separate and that only validity should become a definitional requirement for writing assessment, whereas the centrality of reliability becomes a function of the assessment context (mainly of its purpose and the availability of resources).

Although this proposal may fuel a discussion of the definitional aspects of the validity/reliability controversy, several issues remain unattended. Firstly, there is the question of how to articulate local and global writing assessment when they are both valid for different reasons. Secondly, there is a need to establish when and to what degree reliability estimations may be central to specific contexts. Thirdly, we need to consider the implications of using sources other than reliability estimations (such as teachers' subjective observations of individual characteristics of students/writers) as indicators of validity, and their interpretation by non-specialist stakeholders. Finally, the main implication of this proposal is more contentious: the question is whether constructs can be modified so radically when translated from one discipline to another. Thus a focus on the role of reliability in writing assessment may benefit considerably from closer consideration of the articulation of psychometric assessment constructs and the actual nature of writing events and their circumstances.

References

ASA [American Statistical Association] (2008). *What Is Statistics? What Do Statisticians Do?* Retrieved 8 April 2008 from http://www.amstat.org/Careers/index.cfm?fuseaction=whatisstatistics.

Bodoff, D. (2008). Test theory for evaluating reliability of IR test collections. *Information Processing & Management 44*, 1117-1145.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies 52*, 29-46.

Brown, G.T., Glasswell, K., & Harl, D. (2004). Accuracy in the scoring of writing: studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing 9*, 105-121.

Casanave, C.P. (2004). *Controversies in Second Language Writing: Dilemmas and Decisions in Research and Instruction.* Michigan: The University of Michigan Press.

Davies, A. (1988). Operationalizing uncertainty in language testing: an argument in favour of content validity. *Language Testing 5*, 32-48.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing 25*, 155-185.

Edgeworth, F.Y. (1890). The element of chance in competitive examination. *Journal of the Royal Statistics Society 53,* 460-475.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing 24,* 37-64.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement 58,* 357-381.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In Kroll, B. (ed.) *Exploring the Dynamics of Second Language Writing*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing 12,* 1-9.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist 27,* 353-383.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing 2,* 141-154.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: an empirical study of an analytic scoring rubric. *Applied Measurement in Education 13,* 121-38.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication 18,* 229-249.

Krzanowski, W. J., & Woods, A. J. (1984). Statistical aspects of reliability in language testing. *Language Testing 1,* 1-20.

McKay, P. 2007. *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

Oller, J.W. (1979). Explaining the reliable variance in tests: the validation problem. In Brière, E.J., & Butler Hinofotis, F. (eds), *Concepts in Language Testing: Some Recent Studies*. Washington, D.C: TESOL.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: an empirical study of a holistic rubric. *Assessing Writing 7,* 143-64.

Rudner, L.M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation 3*. Retrieved 9 April 2008 from http://pareonline.net/getvn.asp?v=3&n=3.

Rudner, L.M., & Schafer, W.D. (2001). *Reliability*. ERIC Database. Retrieved 9 April 2008 from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED458213.

Shrout, E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater relibility. *Psychological Bulleting 82,* 420-428.

Skehan, P. (1984). Issues in the testing of English for Specific Purposes. *Language Testing 1,* 202-20.

Slomp, D.H., & Fuite, J. (2004). Following Phaedrus: alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing 9,* 190-207.

Stansfield, C.W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing 5,* 160-186.

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation 9.* Retrieved 10 April 2008 from http://pareonline.net/getvn.asp?v=9&n=4.

Steyer, R. [online]. *Classical (Psychometric) Test Theory*. Jena: Friedrich-Schiller-Universität. Retrieved 10 April 2008 from http://www.metheval.uni-jena.de/materialien/publikationen/ctt.pdf.

Traub, R.E., & Rowley, G.L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement 4,* 517-545.

Underwood, T., & Murphy, S. (1998). Interrater reliability in a California middle school English/Language Arts portfolio assessment program. *Assessing Writing 5,* 201-230.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing 15,* 263-287.

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wiberg, M. (2004). Classical Test Theory vs. Item Response Theory: an evaluation of the theory test in the Swedish driving license test. *Educational Measurement 50,* 1-27.

Daniel Muñoz has been an academic of the Department of Linguistics at the Universidad de Chile since 1996 and a member of its Center for Cognitive Studies since 2001. He holds a BA and an MA degree in English Linguistics. His practice has been mainly concerned with Applied Linguistics, EAP and ESP. As part of his PhD programme at the University of Reading, he is conducting a study which involves the monitoring of the EFL academic writing proficiency development of a group of Chilean undergraduate learners over a one-year period. Besides his general interest in SLA, Applied Linguistics and Cognitive Sciences, he is currently interested in several aspects of writing development, such as language gain measurement and the relation between texts, developmental patterns, writer's individual differences and their instructional contexts. Email: damunoz@uchile.cl.