UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

http://revistas.um.es/ijes

Limited aspects of reality: Frames of reference in language assessment

GLENN FULCHER & AGNETA SVALBERG*
University of Leicester

Received: 26 March 2013 / Accepted: 23 May 2013

ABSTRACT
Language testers operate within two frames of reference: norm-referenced (NRT) and criterion-referenced testing (CRT). The former underpins the world of large-scale standardized testing that prioritizes variability and comparison. The latter supports substantive score meaning in formative and domain specific assessment. Some claim that the criterion-referenced enterprise is dead, save its legacy in score reporting (Davidson, 2012, p. 198). We argue that announcing the demise of CRT is premature. But we do acknowledge that what now passes as CRT is in fact not criterion-referenced, but is based upon a corruption of the original meaning of "criterion" as domain-specific performance. This distortion took place when NRT co-opted the term "standard" to serve as a rationale for the measurement enterprise of establishing cut-scores to retrofit NR tests with meaning derived from external scales. The true heirs of the CRT movement are researchers who base test design in the careful analysis of construct and content in domain specific communication.

KEYWORDS: language testing, criterion-referencing, norm-referencing, domain description, specific purpose testing, scoring criteria, standard setting

RESUMEN
Quienes evalúan el aprendizaje de lenguas operan con dos marcos de referencia: la evaluación basada en la norma (EN) y la evaluación basada en criterios (EC). La primera subyace a la evaluación estandarizada, que prioriza la variabilidad y la comparación, mientras que la segunda fundamenta el significado de los resultados de la evaluación formativa en ámbitos específicos. Hay quienes afirman que la evaluación basada en criterios ha llegado a su fin, dejando como único legado el modo en que se comunican sus resultados (Davidson, 2012: 198). En este artículo defendemos que anunciar la defunción de la EC es prematuro. Sí admitimos, sin embargo, que lo que actualmente se considera EC de hecho no lo es, sino que parte de una corrupción del significado original de "criterio" como actuación relativa a un determinado ámbito. Esta distorsión tuvo lugar cuando la EN se apropió del término "estándar" como sustento teórico para el establecimiento de notas de corte en la actualización de exámenes basados en la norma cuyo significado se extrae de escalas de evaluación externas. Los verdaderos herederos de la EC son los investigadores que basan el diseño de exámenes en un escrupuloso análisis del constructo y de los contenidos de la comunicación específica de cada ámbito.

PALABRAS CLAVE: evaluación de lenguas, evaluación basada en criterios, evaluación basada en la norma, evaluación con fines específicos, criterios de evaluación, definición de estándares

Address for correspondence: Glenn Fulcher, School of Education, University of Leicester, 21 University Road, Leicester LE1 7RF, United Kingdom. Email: gf39@leicester.ac.uk Agneta Svalberg, School of Education, University of Leicester, 21 University Road, Leicester LE1 7RF, United Kingdom. Email: amls2@le.ac.uk

Print ISSN: 1578-7044; Online ISSN: 1989-6131

1. DEFINITIONS

The term "criterion-referenced testing" was first used by Glaser in a series of publications in the early 1960s. The term was used to distinguish a newly conceptualised frame of reference from the existing "norm-referenced testing" (Glaser, 1963). As Glaser (1994a: 9) later put it, "…there was a need for development of proficiency instruments which assessed performance, not in terms of how an individual compared with other individuals, but with respect to how adequately he or she had attained the level of competence required for system operation." The machinery of norm-referenced testing had been evolving since the mid-19th Century, which Edgeworth (1888:626) correctly described as "a species of sortition". Tests were the tools society had designed to rank order individuals for the purpose of decision making, usually for employment or certification. The technology that made test use possible was the curve of normal distribution. The ability of test designers to create items with maximum variance and high discrimination spread the test-taking population out in such a way that the position of any individual could be compared with the proportion of the population gaining a lower or higher score (Fulcher, 2010: 35-42).

Norm-referenced tests and the interpretation of NRT scores are premised on the twin concepts of sortition and comparison, through procedures that establish the relative position of each member of the population. This requires levels of score variability in the test taking population not required in CRT (Popham & Husek, 1964), whereas the central insight of CRT is "…referencing a test to descriptions of achievement or proficiency" (Glaser, 1994a:10). Glaser (1963: 519) says "What I shall call criterion-referenced measures depend upon an absolute standard of quality, while what I term norm-referenced measures depend upon a relative standard." The word "standard" here is problematic, given its current use. Glaser used it interchangeably with the term "criterion", by which he meant actual performance in real-world activities. It is not coincidental that the first reference to criterion-referenced testing was in the context of systems operation, where judgments had to be made about the readiness of trainees to operate machinery or conduct specific tasks within a complex process (Glaser and Klaus, 1962). The problem that CRT addressed was whether test takers could do things, rather than their relative standing in a population.

The meaning of the CRT test score is therefore absolute, in that it indicates whether the test taker meets the criterion, and perhaps how well it is met. The notion of item/task discrimination is still relevant, but it pertains to distinguishing between those who have genuinely met the criterion and those who have not. However, item/test variance becomes largely irrelevant (Millman & Popham, 1974). In educational contexts this is particularly pertinent, as it is expected that in successful curriculum-related achievement tests the majority of students will attain the educational goals that are set, and being taught.

While the definitions of criterion-referenced assessment in the literature differ from source to source (Brown and Hudson, 2002:3-4), the central tenet is clear. As Millman (1972:

Print ISSN: 1578-7044; Online ISSN: 1989-6131

278) puts it, "Although test experts do not agree on a single definition of criterion referenced tests, all variants have in common their emphasis, in interpretation, on what a child can do relative to the subject matter of the test."

2. THE CORRUPTION CRITERION

Establishing a cut score on a test was not part of the original concept of CRT (Glaser, 1994a: 10). It was Mager (1962) who first introduced the notion of cut scores for mastery (see discussion in Glass, 2011), after which the term "criterion" was reinterpreted as a "cut score" on a test that represented having achieved the required "standard". This corruption of the meaning of a term is critical to the move away from the core goals of CRT towards modern "standards-based" assessment, with its external scales and accountability agenda (Hudson, 2012). Instead of trying to define successful performance in a specified domain, attention shifted to the psychometric problem of establishing cut-scores. For example, in their discussion of CRT Hambleton and Novick (1973) focus entirely on cut-scores, and state that the entire problem of CRT is to establish that any observed score on a test is really greater than the cut-score. Such a radical reinterpretation could only have been made by psychometricians from an NRT frame of reference, with little interest in content. In applied linguistics – and *language* testing – the original meaning of "criterion" is retained. Bachman (1990:75), for instance, clearly states:

> It is important to point out that it is this *level of ability* or *domain of content* that constitutes the criterion, and *not* the setting of a cut-off score for making decisions. The definition of a criterion level or domain and the setting of a cut-off score for a given decision are two quite distinct issues. It is quite possible to develop and use a CR test without explicitly setting a cut-off score. (italics in the original).

The educational measurement literature is replete with references to the use of CRTs to assess whether individuals have met "instructional objectives", on the assumption that these are easily defined and listed. Typical of these myopic views of education is this quotation from Hambleton and Novick (1973: 160) "…test information is usually used immediately to evaluate the student's mastery of the instructional objectives covered in the test, so as to locate him appropriately for his next instruction." This is primarily because measurement experts wish to describe mastery as the proportion of items a test taker can answer correctly from the Universe of items that defines a learning objective. This may be a useful fiction for statisticians, but has little relevance for the language teacher. Horne (1984: 155) rightly argues that a purely psychometric approach focused on cut-scores "masks the quality of response". Glass (2011: 233-234) is therefore correct in his view that the conflation of meaning between the two terms "criterion" and "standard" was driven by the interests of psychometrics: "The

Print ISSN: 1578-7044; Online ISSN: 1989-6131

evolution of the meaning of 'criterion' in criterion-referenced tests is, in fact, a case study in confusion and corruption of meaning" (Glass, 2011: 34).

In much modern language testing this corruption has been uncritically accepted. It is widely assumed that particular scores on tests act as "cuts" between levels of mastery that are defined by external "criteria" or "standards". The external standards are usually stated in terms of behavioural objectives. These may be educational taxonomies (Marzano & Kendall, 2007) or language function taxonomies (Council of Europe, 2001); but they remain abstract statements of "standards" that may or may not be relevant to a particular decision context. Unfortunately, the use of such external standards has become common practice in CRT (Berk, 1980: 5), reducing the problem to conducting standard-setting (or "linking/mapping") studies, between test scores and the relevant external standard (e.g., establishing a cut-score on an existing test to describe having achieved level B2, because B2 is the level required for University entry by political mandate; see Fulcher, 2010: 244-248). The corruption is so thorough that even when language specialists call for increased attention to content validity and domain representativeness, the recourse is to a-priori or measurement driven models, rather than frameworks established on the basis of criterion definition (Chalhoub-Deville, 2009; Fulcher *et al.*, 2011).

The assessment tools that we get as a result turn out to be what Hambleton *et al.* (1978:3) refer to (approvingly) as "objectives-referenced tests", in that they do not relate at all to a particular domain of inference, but to abstract lists of objectives to be attained by learners. Hambleton *et al.* treat all criterion-referenced tests as objectives-referenced, and argue that norm-referenced tests may be given criterion-referenced meaning if specific test items can be matched to individual objectives. This has generated a significant amount of psychometric research that aims to retrofit norm-referenced language tests with criterion-referenced meaning, from rule-space methodology (Tatsuoka and Buck, 1998) to fusion theory (Jang, 2009), none of which has been successful when dealing with complex constructs such as language competence (Alderson, 2010). Providing substantive meaning to test scores requires validation by design, rather than validation by retrofit (Fulcher, 2013a). This assumes that we are able to provide descriptions of the domains to which test scores make predictions, and which provide the data for test design.

3. DOMAIN DESCRIPTION

It has frequently been suggested that there is no strict line to be drawn between NRT and CRT (Bachman, 1990:76), as CR tests may be normed using test takers who vary in ability with regard to the criterion, and NR tests may be retrofitted with CRT interpretations. Even advocates of CRT such as Popham (1976: 593) argue against a "toss the rascals out" strategy,

Print ISSN: 1578-7044; Online ISSN: 1989-6131

in favour of incorporating normative data when "how well" questions are asked in addition to "what can they do?"

However, this merger cannot be pushed too far. For scores to carry CR meaning test specifications must incorporate a description of the theory or performance data that supports rich score inferences (Davidson, 2012). It may be the case that some NRT-type activities go on during the process of CRT domain description. The applied linguistic work to describe what language is used to perform domain-related tasks, the discourse structure of interactions, communication strategies and pragmatics, is complex. It may require studying the performance of participants known to be successful, borderline, and weak. One example is the analysis of the performance of potential air traffic controllers, with the aim of making judgments regarding ability to perform tasks in a high stakes environment. But the purpose of the comparison is not to produce a test that rank orders. The purpose is to define minimum competence in relation to specific tasks, such that false positive results are minimized. Unlike NRT, the CRT approach has grown out of work-based assessment. Yerkes (1920: 382-385) describes the process of developing the army "trade tests" during the First World War, in which the first step was an analysis of the tasks carried out in specific jobs in order to define "…the elements of skill and information and judgment which combine to constitute real proficiency" (ibid., 382). Interviews were also carried out with expert practitioners to discover what levels of skill were needed for successful performance. This domain analysis fed directly into the creation of prototype tests that were piloted in group difference studies (Fulcher, 2012:380-381). There is evidence that a similar process was followed to develop the individual English test for conscripts whose first language was not English (Yerkes, 1921:355), although the group (NRT) test was not based on an analysis of military language. Rather, the score ranges that approximated to certain levels of performance on the individual test were established using correlational techniques (an early version of today's "mapping/linking" exercises). It is therefore not surprising that Glaser first considered the notion of CRT in the context of workplace assessment, with specific reference to human-machine interface and efficiency (Glaser & Klaus, 1962). They argued:

> A machinist can be categorized as an apprentice, a journeyman, or a master at his trade. The specific behaviors implied by each of these levels of proficiency can be identified and used to describe the specific tasks an individual must be capable of performing before he achieves one of these skill levels. It is in this sense that measures of proficiency can be criterion-referenced …. Measures which assess performance in terms of a criterion standard thus provide information as to the degree of competence attained which is independent of the performance of others (ibid., 422).

From this quotation is can clearly be seen that CRT assumes a proficiency continuum that is broken down into meaningful categories in terms of what a person can do. What becomes critical is the identification of those behaviours which mark minimum proficiency for the successful completion of tasks in the real world. Horne (1984:158) correctly identifies

Print ISSN: 1578-7044; Online ISSN: 1989-6131

the description of the continuum of learning and the identification of the level of minimal proficiency for the assessment context as the major challenge of CRT (although the smaller the number of categories used, the better in CRT).

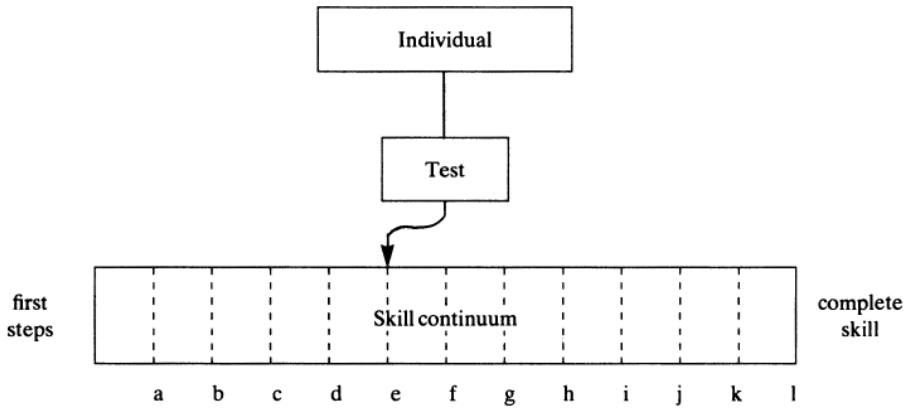

Figure 1. Horne's CRT Model

In language testing, domain definition is somewhat more complex, for even in highly specific domains the range of language functions and tasks can be very large. However, it is not impossible. Discourse analysis may be used to delineate the constructs that are required to complete key domain tasks (Fulcher *et al.* 2011), or the constructs may be operationalized through the careful development of theory (Svalberg, 2009). Sampling tasks from the domain then becomes an empirical matter of prototyping test tasks to discover which options successfully elicit performances/responses that can be scored using data-driven rating instruments. These observed scores are much more likely to be indicative of a "domain score" (Haladyna and Roid, 1983), which has generalizable meaning beyond the immediate context of assessment.

4. USES OF CR TESTS

CR Tests are useful in contexts where it is not important to discriminate between individuals for selection purposes. Their primary utility lies in assessing learning, particularly as achievement tests; providing individualized diagnostic learning; and in assessing language for specific purposes, primarily where minimum competency needs to be established for working environments.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

4.1. Assessing Learning

CR tests are ideally suited for use in classroom assessment, and CRT the theoretical basis for the Assessment for Learning movement, and Dynamic Assessment (Fulcher, 2010:79-81). While comparisons between groups who differ on the criterion provide strong validation evidence, it is not a requirement that such variability should occur when the test is taken. That is, if learning has taken place, we both expect and hope that all test-takers will do well. Indeed, it could be argued that CRT is based on a fundamentally different philosophy from NRT, where by definition 50% of the population will be below average. For in CRT everyone can in principle meet the criterion and be classed as successful. It is therefore not at all surprising that CRT appeals to educators (Hambleton, 1994:22).

In a learning context, the primary purpose is to measure achievement. This is done through specifying carefully what skills and abilities contribute to mastery. Glaser (1994b) summarizes group difference studies that were conducted in order to investigate how masters and novices differed in their responses to criterion-referenced test tasks in a general educational setting. He suggests that masters have the following characteristics:

(a) Structured coherent knowledge: the ability to store chunks of information to which they relate new information.
(b) Proceduralized knowledge: ability to apply knowledge to problem solving rather than just recalling it.
(c) Skilled memory and automaticity: the knowledge base is accessed automatically, leaving memory free to work on new problems.
(d) Effective problem representation: the ability to build a model of the problem from which inferences and solutions may be formed.
(e) Self-regulatory skills: the ability to monitor problem solving and question weak or improbable solutions.

In a language context, Fulcher et al. (2011) argue that learners who function successfully in service encounter environments are able to:

(a) Realize the discourse structure of the service encounter.
(b) Manage communication boundaries in unequal power relationships.
(c) Establish rapport through the use of appropriate pragmatic and strategic language uses.

The linguistic realizations are enumerated, and examples from performance data provided. It is this careful definition that leads to diagnostic usefulness in assessment. Indeed, the role of corrective feedback in learning processes that is central to the Assessment for Learning movement was recognized early on as one of the main uses of information from CRTs. Block (1971:294) referred to this as “optimal feedback/correction” that enhances

Print ISSN: 1578-7044; Online ISSN: 1989-6131

learning through an increase in language awareness (Svalberg, 2007). Feedback in classroom assessment therefore may not involve grades, which are for rank ordering, but statements of achievement and what is to be learned next to move beyond the current level of proficiency; or the use of checklists that mark off the acquisition of contributory skills to successful domain performance (Millman, 1972: 280).

4.2. Language for Specific Purpose Assessment

CRT is especially relevant when developing language tests for specific purposes, especially those designed to certify readiness for practice in fields such as aviation, medicine, or engineering (Douglas, 2000: 15-16). The logic of CRT is both simple and compelling. A language test for nurses should be grounded in an analysis of the tasks undertaken by nurses, the communication that takes place in those tasks, and the linguistic realizations of the specific communicative purposes. It is not appropriate to use a NR test constructed for some other purpose, using evidence from a standard-setting procedure to retrofit that test with additional meaning through the imposition of a cut-score for nurses. Such approaches violate the principles of validation, and are likely to attract litigation when poor decisions are made (Fulcher, 2013b). CR meaning must be introduced during the test design and construction phase, not as a post-hoc activity.

5. TEST DESIGN AND CONSTRUCTION

5.1. Selection of Test Items

In NRT items are piloted and selected for their efficiency in discriminating between individual test takers. This is done using item facility and discrimination indices. In CRT this is undesirable. Rather, tasks or items are selected on the basis of whether they discriminate between groups of learners who are masters and those who are not (group difference studies), or between the same learners before and after a teaching (intervention studies). The aim is to maximize the sensitivity of the items to differences between successful and unsuccessful language users, or the effects educational interventions. Ideally, most tasks/items should be answered incorrectly prior to an intervention, and correctly following an intervention. The CRT notion of discrimination links the test construction method closely with teaching objectives (Glaser & Klaus, 1962: 427).

5.2. Criteria, Standards, and Scales

> Achievement measurement can be defined as the assessment of criterion behavior; this involves the determination of the characteristics of student performance with respect to specific standards (Glaser, 1983: 519).

Print ISSN: 1578-7044; Online ISSN: 1989-6131

With the warning above regarding the meaning of the term "standard", we can now elaborate on the application of the criterion in practice. A CR test score according to Glaser is to be interpreted as mastery of a particular task from a specified domain, or the achievement of some skill or ability on a learning continuum. We have argued that what distinguishes a CR test from an NR test, despite similarities, is the requirement for a CR test to define and operationalize the criterion.

Glaser and Klaus (1962: 433-434) list rating scales and checklists as appropriate tools for scoring performances on domain relevant tasks. The former specify the learning continuum and a mastery point, while the latter enumerate critical behaviours essential to successful task completion. The earliest attempts to design language rating scales that had an empirically founded link between descriptors and observable spoken discourse required the speakers to be at least rank ordered according to an abstract notion of proficiency prior to the analysis of criteria that would adequately separate them (Fulcher, 1993; 1996). With reference to Fulcher's work, Pollit and Murray (1996) also demonstrated the requirement of rank ordering in order to investigate the criteria used by raters to sort speech samples. Scoring systems that are genuinely criterion referenced are now based either on the direct description of domain specific language production (e.g. Fulcher, *et al.* 2011), or the reports of what raters pay attention to in making decisions (e.g. Turner, 2000). Both of these approaches are data-driven, but only the former engages with the description of performance. Rating schemes that rely on performance analysis are therefore the heirs of CRT theory, and meet the validation challenges set out for CR tests: "The degree of validity in a proficiency measure is a function of the difficulties which exist in identifying, quantifying, selecting, and weighting the behaviors to be assessed" (Glaser & Klaus, 1962: 441).

Recent developments in rating scale theory in Fulcher *et al.* (2011) propose the use of performance decisions trees (PDTs) that embody definitions of successful domain specific performance, and require raters to make decisions about a test-taker's ability to realize a range of linguistic, discourse, and pragmatic abilities deemed critical to that domain. By structuring production features in this way, PDTs also fulfill another important role of CR measurement: the potential for providing diagnostic information. Thus, by designing the scoring and reporting models in relation to domain specific inferences, effect-driven testing is achieved: the explicit process of test design with its effect on learning in mind (see also Millman, 1970).

5.3. Task Types and Content Validation

Psychometricians have generally been unable to break away from using NRT type items in CR tests, because their understanding of a "domain" is limited to an exhaustive list of all possible tasks that make up the domain. For example, Linn (1980) assumes that content validity is achieved through random sampling of items that together exhaust the domain definition. All examples are drawn from very narrow areas of basic mathematics. As soon as

Print ISSN: 1578-7044; Online ISSN: 1989-6131

he turns to reading, the only example that works is a cloze test in which there is an assumption that an ability to complete the blanks correctly is a direct measure of comprehension, and that all possible alternative forms will be construct equivalent. Similarly, the illustrations provided by Haladyna and Roid (1983:271) include "all items testing integer addition in the form, i + j = k" from mathematics, and "pronunciation of the 5000 most frequently occurring words in English-language textbooks". The latter is not problematized at all, either in terms of the rationale for the selection of the textbooks, the number of words, the usefulness of pronunciation for any particular purpose, or the relevance to specific communicative contexts. Glass is surely correct when he observes that in such definitions we have

> …pseudoquantification, a meaningless application of numbers to a question not prepared for quantitative analysis. A teacher, or psychologist, or linguist simply cannot set meaningful standards of performance for activities as imprecisely defined as 'spelling correctly words called out during an examination'. (Glass, 2011: 229)

CRT drew attention to what we have only so far alluded to – the importance of content validation for the description and selection of task types that do not under-represent the domain, and which allow extrapolation of score meaning from performance on test tasks to the external domain. As Cronbach (1970:509) observed, "demands for content validity have suddenly become insistent... for data that describe learners rather than rank them; the art of test construction has so far not coped very well with these demands." What we need is a description of the communicative domain and the range of tasks and genres that define that domain. Glaser and Klaus' understanding of content is much more useful in language testing:

> The adequacy of a proficiency test depends upon the extent to which it satisfactorily samples the universe of behaviors which constitute criterion performance. In this sense, a test instrument is said to have *content validity*; the greater the degree to which the test requires performance representative of the defined universe, the greater is its content validity. (Glaser & Klaus, 196: 435; emphasis in the original).

The re-emergence of content concerns (e.g. Lissitz, 2009) in educational measurement is not coincidental. It echoes the rediscovery in language testing of the intimate connection between context and communication (Fulcher *et al.*, 2011); and while construct validation remains of critical importance, the suppression of content issues that seemed so obvious at the end of the 20th century (see Fulcher, 1999) is no longer such a certainty. In CRT, construct validation requires specifying the abilities and processes involved in the domains to which we wish to predict. For example, Haertel (1985) argues that it is only possible to operationalize the construct of "functional literacy" with reference to a specific domain description (the kinds of texts to be read), the context of reading, and the kinds of responses expected as a

Print ISSN: 1578-7044; Online ISSN: 1989-6131

result of reading these texts. Similarly, with reference to learning to communicate in a military setting, Cartier argued that

> The theory requires that the test environment and circumstances approximate those of the work situation, which, for our students, may be a technical school, a maintenance hangar, an aircraft at 40,000 feet, sometimes even somewhere ten fathoms deep. Those circumstances are pretty hard to duplicate, but it may be possible to set up situations in which the student must understand and respond in English under distractions and psychological pressure (Cartier, 1968: 28).

Specifying performance conditions for task types may be critical in many CR tests, and can push the task designer away from closed response items towards more open ended simulation-type tasks (Glaser, 1994b). However, this raises three problems that are not faced in traditional NR tests.

Firstly, as the tasks become longer and more complex, testing time increases. Further, the number of tasks that can be fitted into this time decreases. There is then pressure to increase test length to achieve content-representativeness, resulting in validation vs. length/cost tensions.

Secondly, we face problems with simulation and the nature of abstraction. Test tasks are always abstractions from reality, and communication in many contexts and performance conditions is not easily replicated. Glaser and Klaus (1962: 432, 459) give the examples of fire fighting and military communication under fire. Figure 2 on the next page illustrates an assessment for the crews of amour platoons of five tanks, who have to coordinate their manoeuvres to defeat an enemy tank platoon operated by instructors. The simulation requires the use of radio operated models on a large diorama. Model tanks fire beams of light which can disable enemy tanks when they are hit. It is not difficult to imagine other contexts in which successful communication is critical, but where replicating the performance conditions in a test would violate health and safety requirements.

Thirdly, we must consider the extent to which it is necessary to include "rare events" in the selection of tasks. These are situations that are highly unlikely, but should they occur, communication failure would be extremely serious. Such tasks may be critical for certification in language for specific purpose situations like the assessment of pilots, whose language needs would be very different in the unlikely event of multiple-engine failure, or mid-air collision.

When considering these problems the primary consideration should be that design decisions should not be driven by cost alone, but by risk estimates associated with false positive results in high-stakes assessment.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

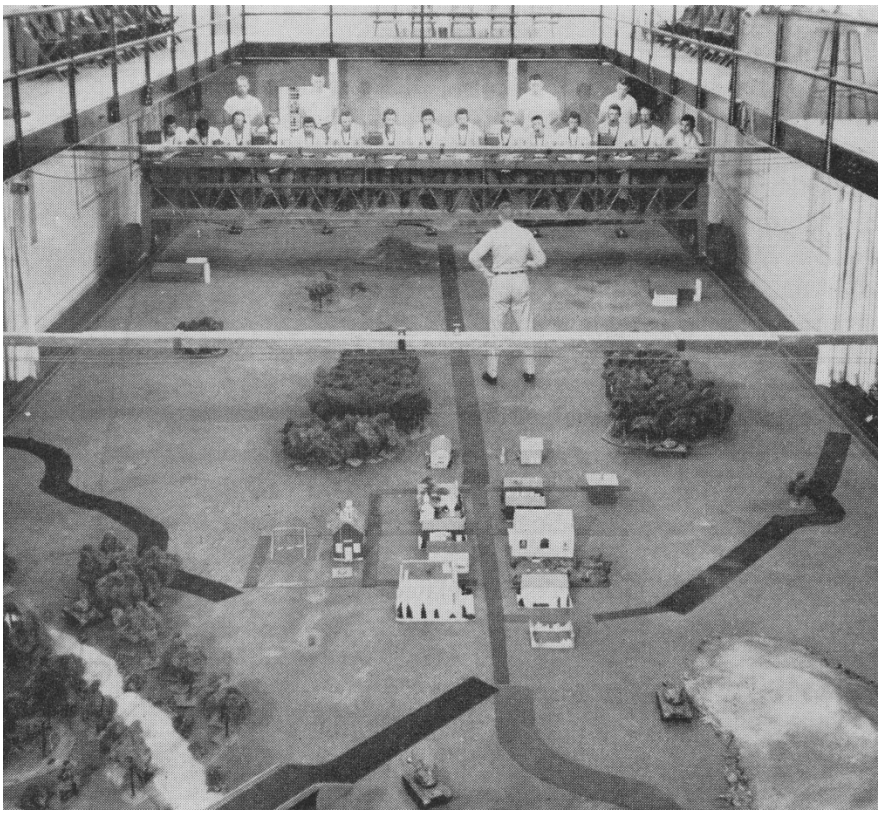

Figure 2. A tank crew assessment (Glaser & Klaus, 1962)

5.4. Test Specifications

The new CRT technology that linked together all the elements discussed above into the test design process was the test specification. Test specifications are the test design blueprint that describes test purpose, the domain of inference, and what each item or task is designed to measure; as generative, iterative, and consensus-based documents, they provide all the details necessary for the generation of task pools and test forms (Davidson, 2012: 198). Popham (1978) provided the first test specification templates that are still in use today (see examples in Brown & Hudson, 2002: 88-95). They have also been used in language testing to achieve "reverse washback" – or the effect of teaching context on the design of classroom tests (Davidson & Lynch, 1994). In assessment for learning the test specification becomes the focus of articulating learning objectives and how they can be formatively assessed. The consensual nature of the process is stressed because it empowers teachers through collaboration and professional development.

The test specification mediates between the domain of interest and test content, which is necessarily an abstraction of the real world domain. It is the explicit statement of what is important about the domain that can be replicated in the test. As such, it provides validation evidence by design, through the notion of item/task-criterion congruence (Popham & Husek,

Print ISSN: 1578-7044; Online ISSN: 1989-6131

1969) on the one hand, and Item/task-specification congruence (or "fit-to-spec"), on the other (Davidson & Lunch, 2002: 44-48).

The success of specifications in CRT has led to their use in all testing enterprises today, thus providing NRTs with some of the trappings of CRT interpretation.

6. TEST PROPERTIES

6.1. Cut Scores and Standard Setting

While the use of cut-scores was not originally part of the CRT enterprise, it does not seem unreasonable that cut-scores are used in some contexts to implement mastery decisions. Cizek and Bunch (2007: 14) place the need for cut scores in the need of society to make proficiency decisions and, "…when we speak of 'setting performance standards' we are…referring to the…concrete activity of deriving cut points along a score scale." If a test has been designed along CRT principles, establishing cut-scores that represent gradations of achieving criterion performance can be useful.

Many standard-setting methods have been developed. The most commonly used in instructional settings are completely arbitrary, including using the median score of the group, or deciding which proportion of correct items short of 100% represents mastery. Many institutions select 60% (or a grade C) as a default, without any explicit rationale other than precedent. A range of more explicit methods have been developed, most of which require the judgment of experts upon the likely success of minimally competent students on the items or tasks on the test (for descriptions see Berk, 1986; Brown, 2013; Cizek & Bunch, 2007; Fulcher, 2010). One method that does not rely on expert judgment is the criterion-groups validation model (Berk, 1976). This requires the administration of a test to a group of known masters and non-masters, perhaps on the basis of instruction. The cut score may then be established at a point separates the two groups most efficiently (see Figure 3), or shifted to minimize either false positives or false negatives depending on the nature of the decisions being made (see Fulcher, 2010: 240). Figure 3 makes it clear that this approach, like all standard setting methods, makes NR assumptions about test-taker distributions.

Figure 3 on the next page also illustrates Cizek and Bunch's (2007: 18) statement that "Standard setting does not seek to find some pre-existing or "true" cutting score that separates real, unique categories on a continuous underlying trait…." Rather, standard-setting activities are compared with following legal processes:

> According to the relevant legal theory, important decisions about a person's life, liberty, or property must involve due process - that is, a process that is clearly articulated in advance, is applied uniformly, and includes an avenue for appeal (Cizek & Bunch, 2007: 15).

Print ISSN: 1578-7044; Online ISSN: 1989-6131

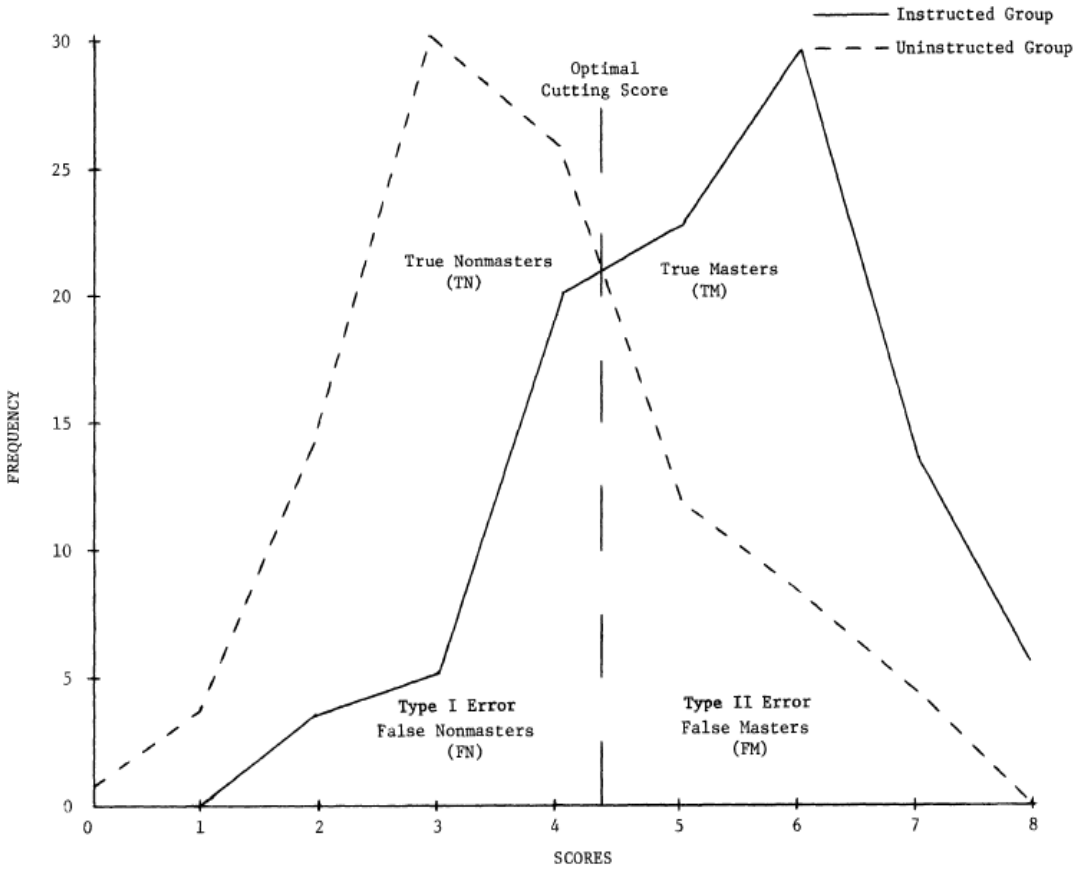

Figure 3. Criterion-groups validation model (Berk, 1976, p. 6)

It is following the procedures as laid down by the standard-setting procedure selected that provides faith in the cut scores, rather than any notion of "truth". This admission of arbitrariness is unique in educational measurement (Ebel, 1971: 287). As Glass (2011: 254) puts it,

> To my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. But arbitrariness is no bogeyman, and one ought not to shrink from a necessary task because it involves arbitrary decisions. However, arbitrary decisions often entail substantial risks of disruption and dislocation. (Glass, 2011: 254).

However, it must be remembered that current standard-setting approaches do make NRT assumptions, and for the most part use fallible judges to arrive at cut-scores. There is little development of CR tests based on analyses of successful performance in specific domains. If applied to CR tests, this arbitrariness would be substantially removed.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

6.2. Reliability and Dependability

Many educational measurement specialists who believe that it is possible to produce sets of items that exhaust a domain, also argue that there must be variability within the domain, so that a particular proportion of sample items answered correctly represents mastery as defined by the cut score. Thus, the argument goes that internal consistency is still relevant to CR tests (Kane, 1986). However, it is more commonly accepted that there is a genuine difference in the meaning of reliability across the two paradigms.

> …the reliability of a criterion-based score can be thought of as the *accuracy of measurement*, while the reliability of a norm-referenced measure can be thought of as the *consistency* of a score. (Glaser & Klaus, 1962: 446; italics in the original).

This "accuracy" is essentially the dependability of the decision that a performance is at mastery level (Davidson, 2012: 198). The most significant threat to dependability is rater/judge variability or bias, but also includes facets of the test method. Approaches to assessing dependability include agreement of mastery classifications, agreement of decisions at cut scores, and the dependability of domain scores (Berk, 1980; Kunnan, 1992). Dependability of mastery classifications are the most useful, and Brown (2013: 4) recommends the use of phi-lambda (φλ) for language testers as a (threshold loss) index, because it "provides an estimate of the degree of score consistency for decisions made with a particular cut score, while taking into account the distances of the scores from the cut score." This dependability index with a CR confidence interval is easily calculated by hand (Brown & Hudson, 2002: 195-198; Fulcher, 2010: 84-86). Two other useful approaches include looking at consistency of classification across alternative forms of an assessment to investigate loss of dependability because of test method facets, and across repeated administrations to estimate test-retest dependability. However, it is rare that institutions which claim to have developed a CR test provide CR dependability statistics. This is because the CRT interpretation is usually at best a veneer.

7. CONCLUSION

What now passes for criterion-referenced assessment in language testing is but a pale reflection of the promise offered by criterion-referencing theory. Establishing cut-scores with reference to descriptive models like the CEFR is illusory, and only serves to subvert validation theory. The current fad for linking and mapping is based upon the corruption of "criterion". Rather, we agree with Cartier (1968: 32) who argued that "Criterion tests insist on actual behaviour…", which returns us to the original motivation of Glaser: to invest scores with more meaning about achievement and likely behaviour in a specified domain beyond the

Print ISSN: 1578-7044; Online ISSN: 1989-6131

test. The majority of language tests that claim to be criterion-referenced are no different from those that Popham criticized in 1978:

> Every one of these misnamed measures was considered criterion-referenced by its developers merely because the developers had gone to some trouble in identifying a test cutoff score below which an examinee's performance would be considered inadequate. None of the tests provided a more incisive description of examinee performance than one would find in a norm-referenced test (Popham, 1978: 92).

Even if the external descriptive scheme is detailed, it still does not make a test criterion-referenced: "Merely hooking up a test to an imprecisely stated objective, even a behavioural one, fails to delimit satisfactorily the behaviors being assessed by the test" (ibid., p. 93). This is particularly the case with external a-priori scales, because they are not criterion definitions but circular constructions: you are a level 2 learner because you can do level 2 tasks, and level 2 tasks are so defined because they can be successfully performed by level 2 learners (Brown & Hudson, 2002: 26). This is not the same thing as describing a CRT domain.

The future of CRT in language testing lies with those who are attempting to carefully define the domains to which inferences are to be made from test scores. With reference to task analysis and the identification of critical performance elements for inclusion in scoring instruments, data-driven approaches continue to make headway (Fulcher, *et al.*, 2011). This work is also being applied beyond the speaking context by other researchers (e.g. Kim, 2010; Knoch, 2009). Significant work is also being undertaken with respect to the definition of performance in academic contexts, which informs versioning strategies for the development of tests of academic English (e.g., Biber, 2006; Rosenfeld *et al.*, 2001). Further applied linguistic description of domain specific communication is urgently called for.

The success of these efforts will be determined by whether they are more successful than retrofitted NR tests in generating scores from which sounder inferences can be made, leading to more dependable domain specific decisions. Validation efforts of new CR tests will centre on their ability to better predict future performance. It is also possible to conduct experimental research using CR tests, using group difference studies with groups who are known to differ on the criterion on test-external grounds; intervention studies are also possible, as CR tests should be more sensitive to instruction than their NR counterparts (Glaser & Klaus, 1962: 448). It is this sensitivity that should make CR tests diagnostically useful (Hambleton, 1994: 25).

Finally, we must recognize that CR tests have grown out of a descriptive tradition that is closely associated with making decisions about readiness to practice in the workplace (Fulcher, 2012). It is therefore likely that CR tests will remain high stakes decision making tools. CRT theory reminds us that despite the statistical machinery of measurement theory which helps us produce sound assessments, language testing remains a social science. At its

Print ISSN: 1578-7044; Online ISSN: 1989-6131

core remains language use in social contexts, and our ability to predict successful communication for the benefit and safety of all stakeholders.

REFERENCES

Alderson, J. C. (2010). Cognitive Diagnosis and Q-Matrices in Language Assessment: A Commentary. *Language Assessment Quarterly, 7*(1), 96-103,

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Berk, R. A. (1980). *Criterion-Referenced Testing: the State of the Art*. Baltimore: John Hopkins University Press.

Berk, R. A. (1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. *Review of Educational Research, 56*(1), 137-172.

Biber, D. (2006). *University Language. A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Block, J. H. (1971). Criterion-Referenced Measurements: Potential. *The School Review, 79*(2), 289-298.

Brown, J. D. (2013). Cut Scores on Language Tests. In Chapelle, C. A. (Ed.), *The Encyclopedia of Applied Linguistics*. London: John Wiley and Sons.

Brown, J. D. and Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

Cartier, F. A. (1968). Criterion-Referenced Testing of Language Skills. *TESOL Quarterly, 2*(1), 27-32.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In Lissitz, R. W. (Ed.) *The Concept of Validity* (pp. 241-263). Charlotte, NC: Information Age Publishing.

Council of Europe. (2001). *Common European framework of reference for language learning and teaching*. Cambridge, UK: Cambridge University Press. Available online at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Cronbach, L. J. (1970). Review of *On the theory of achievement test items. Psychometrika, 35,* 509-511.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Applied Linguistics* (pp. 197-207). London and New York: Routledge.

Davidson, F. & Lynch, B. K. (1994). Criterion-Referenced Test Development: Linking Curricula, Teachers, and Tests. *TESOL Quarterly, 28*(4), 727-743.

Davidson, F. and Lynch, B. K. (2002). *Testcraft. A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT: Yale University Press.

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

Ebel, R. L. (1971). Criterion-Referenced Measurements: Limitations. *The School Review, 79*(2), 282 – 288.

Edgeworth, F. Y. (1888). The Statistics of Examinations. *Journal of the Royal Statistical Society, 51*(3), 599-635.

Fulcher, G. (1993). The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language. Unpublished PhD dissertation. Lancaster University.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (1999). Assessment in English for Academic Purposes: Putting content validity in its place. *Applied Linguistics, 20*(2), 221-236.

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

Fulcher, G. (2012). Scoring Performance Tests. In G. Fulcher & F. Davidson (Eds.) *The Routledge Handbook of Language Testing* (pp. 378-392). London and New York: Routledge.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

Fulcher, G. (2013a). Test design and retrofit. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 5809-5817). Malden MA: Wiley Blackwell.

Fulcher, G. (2013b). Language testing in the dock. In A. J. Kunnan (Ed.), *The Companion to Language Testing*. London: Wiley-Blackwell.

Fulcher, G., Davidson, F. & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing, 28*(1), 5-29.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Glaser, R. (1994a). Criterion-Referenced Tests: Part I. Origins. *Educational Measurement: Issues and Practice, 13*(4), 9-11.

Glaser, R. (1994b). Criterion-Referenced Tests: Part II. Unfinished Business. *Educational Measurement: Issues and Practice, 13*(4), 27-30.

Glaser, R. and Klaus, D (1962). Proficiency Measurement: Assessing Human Performance. In Gagné, R. M. (Ed.) *Psychological Principles in System Development* (pp. 421-427). New York: Holt, Rinehart and Winston.

Glass, G. V. (2011). Standards and Criteria. *Journal of MultiDisciplinary Evaluation, 7*(15), 227-257. (Originally published in 1977 as an Occasional Paper).

Haertel, E. (1985). Construct Validity and Criterion-Referenced Testing. *Review of Educational Research, 55*(1), 23-46.

Haladyna, T. M. and Roid, G. H. (1983). A Comparison of Two Approaches to Criterion-Referenced Test Construction. *Journal of Educational Measurement, 20*(3), 271-282.

Hambleton, R. (1994). The Rise and Fall of Criterion Referenced Measurement? *Educational Measurement: Issues and Practice, 13*(4), 21-26.

Hambleton, R. & Novick, M. R. (1973). Toward an Integration of Theory and Method for Criterion-Referenced Tests. *Journal of Educational Measurement, 10*(3), 159-170.

Hambleton, R., Swaminathan, H., Algina, J. & Coulson, D. B. (1978). Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments. *Review of Educational Research, 48*(1), 1-47.

Horne, S. (1984). Criterion-referenced testing: pedagogical implications. *British Educational Research Journal, 10*(2), 155-173.

Hudson, T. (2012). Standards-based testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Applied Linguistics* (pp. 479 - 494). London and New York: Routledge.

Jang, E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31-73.

Kane, M. (1986). The Role of Reliability in Criterion-Referenced Tests. *Journal of Educational Measurement, 23*(3), 221-224.

Kim, Y-H. (2010). An Argument-Based Validity Inquiry into the Empirically-Derived Descriptor-Based Diagnostic (EDD) Assessment in ESL Academic Writing. University of Toronto, PhD Thesis. Retrieved 18th March 2013 from https://tspace.library.utoronto.ca/bitstream/1807/24786/1/Kim_Youn-Hee_201006_PhD_thesis.pdf.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing, 9*(1), 30-49.

Linn, R. L. (1980). Issues of Validity for Criterion-Referenced Measures. *Applied Psychological Measurement, 4*(4), 547 – 561.

Lissitz, R. W. (2009). *The Concept of Validity*. Charlotte, NC: Information Age Publishing.

Lynch, B. & Davidson, F. (1994). Criterion-Referenced Language Test Development: Linking Curricula, Teachers, and Tests. *TESOL Quarterly, 28*(4), 727-743.

Mager, R. F. (1962). *Preparing Instructional Objectives*. Palo Alto, CA: Feardon Publishers.

Marzano, R. J. & Kendall, J. S. (2007). *The New Taxonomy of Educational Objectives*. Second Edition. Thousand Oaks, CA: Corwin Press.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

Millman, J. (1970). Reporting Student Progress: A Case for a Criterion-Referenced Marking System. *The Phi Delta Kappan, 52*(4), 226-230.

Millman, J. (1972). Criterion Referenced Measurement: An Alternative. *The Reading Teacher, 26*(3), 278-281.

Millman, J. & Popham, J. (1974). The Issue of Item and Test Variance for Criterion-Referenced Tests: A Clarification. *Journal of Educational Measurement, 11*(2), 137-138.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to? In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74-91). Cambridge, England: Cambridge University Press.

Popham, J. (1976). Normative Data for Criterion-Referenced Tests? *The Phi Delta Kappan, 57*(9), 593-594.

Popham, J. (1978). *Criterion Referenced Measurement.* Englewood Cliffs, NJ: Prentice Hall.

Popham, J. (1994). The instructional consequences of criterion referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15-18.

Popham, J. and Husek, T. R. (1964). Implications of Criterion-Referenced Measurement. *Journal of Educational Measurement, 6*(1), 1-9.

Rosenfeld, M., S. Leung, & P. K. Oltman. (2001). *The reading, writing, speaking, and listening tasks important for academic success at undergraduate and graduate levels*. TOEFL Report MS-21. Princeton, NJ: Educational Testing Service.

Simon, G. B. (1969). Comments on "Implications of Criterion-Referenced Measurement. *Journal of Educational Measurement, 6*(4), 259-260.

Svalberg, A. (2007). Language awareness and language learning. *Language Teaching, 40*(4), 287-308.

Svalberg, A. (2009). Engagement with language: interrogating a construct. *Language Awareness, 18*(3-4), 242-258.

Tatsuoka, K. and Buck, G. Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Turner CE (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review, 56*(4), 555–584.

Yerkes, R. M. (1920). What psychology contributed to the war. In R. M. Yerkes (Ed.), *The New World of Science: its development during the war* (pp. 364-389). New York, NY: The Century Co.

Yerkes, R. M. (1921). *Pschological Examining in the United States Army*. Memoirs of the National Academy of Sciences, Vol. XV. Washington, DC: Government Printing Office.

Print ISSN: 1578-7044; Online ISSN: 1989-6131