A Case for Greater Linguistic Input into the Evaluation of ESL Objective Tests

Roseline Abonego Adejare
Department of Language, Arts and Social Science Education, Lagos State University, Ojo, Lagos
Email Address: abonego@yahoo-com

Abstract

This paper made a case for greater linguistic input into the evaluation of ESL Objective Tests in order to enhance the tests' quality. There are no evaluative criteria specifically for ESL Objective Tests, and the general criteria for language test evaluation lack sufficient linguistic input to meet the peculiar demand of ESL Objective Tests. Language test evaluation is not known to have been theorised from the perspective of the linguistic items that constitute them. The paper critically examined existing criteria for evaluating language tests, and adopted validity as an all-embracing, superordinate term for language test evaluation since reliability and practicality are necessary conditions for validity. It proposed three linguistics-informed criteria for establishing the validity of ESL Objective Tests without necessarily jettisoning other sources and types of validity evidence exemplified by scoring validity. The criteria are linguistic content (adequacy of coverage, test items, and distracters), testing techniques (suitability for testing goals, and language and skills areas), and testees' performance (correlation with actual language ability). To mitigate the inadequacies of ESL Objective Tests, the capacity of the tests' linguistic content to realise the defined goals of testing should be prioritised during its determination.

Key Words: Language Test Evaluation ESL Objective Tests Criteria for Validity Linguistic Input Test Validation Linguistics-based Criteria

Introduction

Language tests are conventionally evaluated using three criteria which also double as the primary qualities that a good language test should possess. These are validity, reliability and practicality. As a rule, a good test must measure what it purports to measure. This is known as validity. Furthermore, scores obtained by testees in repeat attempts of the same test and scorers' marking must be consistent for a test to be reliable. Lastly, any good test must be administratively economical and practicable (Lado, 1961; Grieve. 1964; Harris 1969; Bachman & Palmer, 1996). A critical examination of the literature reveals that much of what is done in the name of validity determination of language tests centre around scores obtained as assessment data. Put more straightforwardly, the validity process revolves around the statistical handling of scores in various ways, with little or no attention paid to the actual linguistic data provided by the individual test items themselves. In essence, there is insufficiency of input from linguistics in the whole exercise called language test evaluation, whereas issues surrounding language tests are intrinsically more linguistic than educational and psychological. If a language test measures linguistic behaviour, or an individual's language ability, as Harris (1969) and Bachman and Palmer (1996) respectively stated, it then becomes imperative that a critical aspect of its validation before, during and after its administration should involve sufficient input from linguistics. The need for greater linguistic input into language test validation becomes even more critical where the language variety concerned is ESL and the test in question is ESL Objective Tests.

There are strong reasons why linguistic input should be recognised, acknowledged and adequately utilised in the process of validating ESL Objective Tests. The first is the multiple roles that English plays in Nigeria's national development. It is Nigeria's official language and consequently the language of education, sophisticated commerce, and mass media. English serves as a unifying force by bridging the communication gap among the formally educated members of Nigeria's linguistically heterogeneous community, in addition to being the wider window through which Nigeria views the outside world. It is a second language at both the individual and national levels, and this has implications for its dual role in education as a compulsory subject of study and the sole medium of instruction in secondary schools. Great attention has, therefore, to be paid to its testing. The second reason relates to the objective technique itself, which has been heavily criticised for being overly simplistic, mechanistic, low in linguistic content, and for encouraging cheating and guessing (Adewumi, 1986; Adejare, 1987). Objective tests are arguably the most popular form of tests in English Language in Nigeria today. Apart from constituting the sole test for selecting candidates for secondary and tertiary education, respectively exemplified by the National Common Entrance and the Unified Tertiary Matriculation Examination, they form a substantial part of the Senior Secondary Certificate English Language Examinations, an achievement test that represents the terminal level for the formal testing of English for most Nigerians.

Based on the foregoing background, this paper makes a case for greater linguistic input into the evaluation of ESL Objective Tests to enhance the tests' quality. It reviews existing criteria for language test evaluation and proposes linguistics-informed criteria for use in the validation of ESL Objectives Tests, in addition to the existing statistics-based criteria. There are no evaluative criteria specifically for ESL Objective Tests, and the general criteria for language test evaluation lack sufficient linguistic input to meet the peculiar demand of ESL Objective Tests. Language test evaluation is not known to have been theorised from the perspective of the linguistic items that constitute them.

The two sections after this introduction review different aspects of the existing criteria for language test evaluation. The fourth section articulates the linguistics-informed criteria for establishing the validity of ESL Objective Tests while the fifth and final section concludes the paper.

General criteria for evaluating language tests

It was stated above that reliability, validity and practicality are the three generally acknowledged criteria used in evaluating language tests. Although some scholars make claim to two or more others, this critical review centres on reliability and validity because they are regarded as the essential qualities of language tests (Bachman & Palmer, 1996). Reliability is handled right away in this section while validity is delayed till the next section for reasons bothering on its theoretical affinity with the main argument of this paper. Meanwhile, it is useful to examine those other criteria that lie outside the triad of reliability, validity, and practicality.

Fairness and other criteria

Le and Klein (2002) identified fairness in addition to reliability and validity as the three main technical criteria for evaluating tests. Fairness as a criterion literally means that the test must be fair to all testees. In other words, the test and test items must be bias-free, "examinees must have comparable opportunities to demonstrate their knowledge and skills", and they must have "sufficient opportunity to learn the tested material" (p. 68). In this regard, fairness partially resembles the practicality criterion, which prescribes that the conditions under which the test is written must not be unfavourable in order to not impinge upon validity. According to Bachman and Palmer (1996), practicality "pertains primarily to the ways in which the test will be implemented, and to a large degree, whether it will be developed and used at all" (p.35). Thus, a test will be impractical if the resources (human, material, and time) for its implementation are unavailable or inadequate. This possibly explains why the WAEC discontinued with the Spoken English component of the Senior School Certificate English Language Examinations. So, the practicality criterion concerns test implementability. Considerations for practicality are crucial for the test administration stage of test development, but are not pertinent for post-test administration evaluation.

In *The Guidelines for the Assessment of English Language Learners*, Educational Testing Service (ETS) (2009) listed three criteria for test validity as follows: test purpose, construct definition, and assessment specifications. According to the document, "the purpose of a test must be clear in order for valid interpretations to be made on the basis of test scores". Similarly, the construct the test intends to measure should be precisely and explicitly defined, and assessment specifications should, among others, include test content and domain of knowledge (pp.8-9). On their part, Aryadoust, Zakaria, Lim, and Chen (2020) posited that examining the validity of tests would minimally require reliability and the psychometric evidence, which would show that both the tasks and the construct measured functioned properly. They did not however identify what the other criteria would be.

Reliability as an essential criterion

The reliability criterion is mainly concerned about test scores and their consistency. A test is adjudged reliable if the scores obtained remain the same irrespective of differences in testing situation and marker personality. The numerous definitions of reliability attest to these statements' veracity. For instance, Bachman and Palmer (1996) defined reliability simply as "consistency of measurement", and added that "a reliable test score will be consistent across different characteristics of the testing situation". In other words, reliability is "a function of the consistency of scores from one set of test and test tasks to another" (pp.19-20). Le and Klein (2002) defined **r**eliability as "the degree to which a test's scores are free from any type of chance effects" and, like Bachman and Palmer (1996, p.38), described it as "a characteristic of the scores on a test as distinct from a characteristic of its questions" (p.52). ETS (2009) similarly defined reliability as "the extent to which scores obtained on a specific form of assessment can be generalized to scores on other forms of assessments, administered at other times, or possibly scored by some other rater (s)" (p.10). Aryadoust et al. (2020) did not define reliability but cited Field (2018) and Green and Salkind (2014), who respectively referred to it as "the evidence that the measurement is precise or has low error of measurement" and "its output is reproducible across occasions, raters, and test forms" (p.2).

According to Le and Klein (2002, p.56), there are three ways in which reliability can be established. These are (1) giving alternate forms (of the test), (2) test-retest (where the same test is taken twice), (3) and internal consistency (which correlates different test questions using a single form of the test, e.g., even v odd numbers). All three criteria are however statistics-oriented, and this is inadequate. Granted that statistics cannot be divorced from reliability because it concerns test scores, statistical tools should not be the sole parameter for establishing the reliability of language tests. Earlier scholars in language testing had recognised the absolute necessity for a linguistic criterion, when they stated that reliability can be empirically established by comparing scores against testees' actual ability to use the skills tested, in addition to already established norms (Lado,1961; Grieve,1964; Harris, 1969). Since a reliable test is a valid test, it follows that scores obtained in an objective test on the Simple Present Tense, for instance, must match testees' actual ability to use the two morphological variants of the tense form in oral and written composition on say, "What I do every day". The test will otherwise be

considered invalid. After all, "test scores are only useful to the extent that they accurately reflect the students' mastery of the domain the test is designed to measure" (Le & Klein, 2002, pp.51-52).

Certain factors may affect the reliability of test scores. Again, Le and Klein (2002, pp.52-55) identified four thus: (a) item sampling (in which a student's score may vary depending on the particular sample of questions that are asked about a topic and the nature of the student's knowledge about the subject), (b) transitory factors (such as illness, anxiety, time of day and time limits, and room lighting), (c) rater agreement (Well-constructed and field-tested scoring guide and extensive rater training will yield a high degree of rater agreement in essay tests), and (d) test length and format (Longer tests with more questions give more reliable scores). They specifically noted that Multiple Choice Question tests cannot assess certain types of knowledge and skills, which explains why scoring reliability is not the sole criterion for evaluating the quality of tests. These views are consistent with Bachman and Palmer's (1996), who explained that inconsistency, attributable to "factors other than the construct we want to measure" (p.38), implies variation in test scores. This shows that a reliable test does not necessarily translate to a valid test, which is why a non-scores-based approach to validity determination should be sought and adopted to complement existing criteria.

The validity criterion

This review of the validity criterion covers its conceptualisation, relevance, types, and investigation in that order.

What is validity?

The title of this subsection is deliberately couched in an interrogative clause. To provide answers to the question "what is validity?", therefore, it is necessary to start by examining its definitions. Ten of these are listed as in (a)-(j).

(a) "The problem of validity is that of whether a test really measures what it purports to measure." (Kelly, 1927, p.14, cited in Weir, 2005, p.12)

(b) "Does a test measure what it is supposed to measure? If it does, it is said to be valid." (Lado,1961, p.321)

(c) "Every time an educator asks 'but what does the instrument really measure' he is calling information on construct validity." (Cronbach, 1971, p. 463; cited in Weir, 2005, p.12)

(d) "Validity means the extent to which the scores on a test provide an accurate information for the decision that will be based on those scores." (Cronbach,1971; Messick, 1987; cited in Le & Klein, 2002, p.60)

(e) "Validity is not an inherent characteristic of a test. It is a characteristic of the decisions that are based on the test scores and the subsequent uses of those scores." (Cronbach, 1988; cited in Le & Klein, 2002, p.60)

(f) "The term construct validity is therefore used to refer to the extent to which we interpret a given test score as an indicator of the ability(ies), or construct (s), we want to measure." (Bachman & Palmer, 1996, p. 21)

(g) "The term validity is often used to encompass the reasonableness of both score-based inferences and the effect of a given use of a test, which are then referred to as consequential validity." (Le & Klein, 2002, p.60)

(h) "Validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e. test scores, which are an accurate representation of a candidate's level of knowledge or skills. In this revision, validity resides in the scores of a particular administration of a test rather than in the test *per se*". (Weir, 2005, p. 12)

(i) "the extent to which a test measures what it claims to measure" (ETS ,2009, p.2)

(j) "A good test must be valid. This means that it should measure what it is supposed to measure; it should predict some criterion, that is, have a correlation with the criterion." (Sheeba, 2020, p. 5)

Two pertinent observations have to be made at this point. The first is the recurrence of the verb *measure* in sixty percent of the definitions (precisely a-c, e, f, i, &j), which makes it significant. Because *measure* is the base lexeme for *measurement*, a term that features prominently in language assessment discourse, it becomes useful to explore both words further in order to explicate the concept validity. The OED defined measure as "to judge the extent or value of something" and measurement as "the activity or process of measuring something" (p.726). Although no definition was found for measure with reference to language assessment, there was one in respect of measurement. According to Aryadoust et al. (2020), who cited two sources, "measurement is the process of (i) construct representation, (ii) construct operationalization, (iii) data collection via behaviour elicitation (Norris & Ortega, 2002, p.720), (iv) data analysis to generate evidence, and (v) the employment of that evidence to draw theory-based conclusions (Messick, 1989, 1996)" (p.3) (See also Sheeba, 2020, p.4). Thus, validity is intrinsically associated with the measurement of *construct* (See c &f), which refers to "the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task" (Bachman & Palmer, 1996. p.21). It also refers to the "skill/proficiency an assessment is intended to measure" (ETS, 2009, p.4). Validity is inherently concerned with test content which, in the case of language tests, is linguistic. It is particular about whether or not a test "measures what it purports to measure", and thus probes into whether the knowledge and skill areas claimed to be tested are what the test actually turned out to test. Thus, a test is invalid

if its contents run contrary to what it sets out to test; otherwise, it is valid. This is traditionally known as content validity.

The second observation is about the plural noun *scores* (their consistency, accuracy, and capacity to truly reflect testees' actual ability in the knowledge areas and skills tested), which occurred in forty percent of the definitions (precisely d, e, g & h). Thus, for ETS (2009), "it is critical to consider the degree to which interpretations of their test scores are a valid reflection of the skill or proficiency that an assessment is intended to measure" (p.2)**.** For a test to be valid, therefore, scores obtained must provide accurate information on which far-reaching decisions about their usage can be based. This relates to reliability. So, a test is valid if the scores obtained and the interpretations resulting therefrom are reliable. It is important to remark that not one definition makes reference to the conditions of test administration as a criterion for validity and that there is a noticeable shift in emphasis from test content to test scores over the decades. With the singular exception of (i) and (j), which appear to have merely echoed earlier definitions, all definitions from Year 2000 restrict validity to reliability of test scores. (Definition (d) does not inhabit this time frame, but for its year of citing.)) This provides further justification for the current argument.

Relevance of validity

Two studies serve to demonstrate the relevance of validity, even though they are not on language testing. The first is on the Multiple-Choice Objective test, about which Teemant (2013) reported that ESL-using students of History preferred the test because it demonstrated content knowledge. This raises the question whether their preference is linked to the fact that the testing technique makes no demand on the productive skills of speaking and writing and very little demand on the receptive skills (reading only). The second is on Language Assessment Literacy (LAL), seen as "a repertoire of competence or skills and specific levels of knowledge…that enable an individual to understand, assess, construct language tests and analyse and interpret test data effectively" (Osidak, Drahinda & Kvasova, 2021, p.2, citing Bohn & Taylor, 2021). Bustamante (2022) investigated the LAL levels of Middle School English Teachers in Mexico, and found that one area in which teachers desired knowledge was "developing and interpreting evaluation criteria" (p.14). These reports further justify why validation studies must be carried out.

Types and sources of validity

Lado (1961) and Grieve (1964) listed what might traditionally be regarded as the sources and types of validity, when they separately stated that validity depends largely on linguistic content, the situation and technique used in testing the content, and the absence of factors extraneous to linguistics. However, recent developments have seen new types and terminologies emerge. For instance, Weir (2005) identified (1) construct validity (described as "a function of the interaction of theory-based validity and context validity), (2) scoring validity, (3) criterion-related validity (where a test is measured against other external measures of the construct), and (4) consequential validity (the argument that test scores have social values and social consequences and that an account of their impact should extend beyond the test taker to include other stakeholders such as the school, teachers, and curriculum (pp.14-15). It was on account of this that Saglam, Lidice and Tsagari (2022) studied aspects of consequential validity of integrated language proficiency tests in a Turkey university setting, and reported that its absence resulted in the narrowing of the curriculum to tested areas only. The Standards for Educational and Psychological Testing (American Educational Research Association) and National Council on Measurement in Education (2014) recognised test content, response processes, internal structure, relationship to other variables, and testing consequences as sources and types of validity, according to Hoeve (2022, p.3). For the purpose of this paper two types of validity shall be recognised, namely content validity and scoring validity.

Content validity

When test developers show concern about whether the content of the test represents exactly what they intend to test (nothing more, nothing less), they are dealing with content validity or validating the test's content. Weir (2005) used the nomenclature *context validity* instead of content validity, and it means "the extent to which the choice of tasks in a test is representative of the larger universe of which the test is assumed to be a sample" (p.19). Content validity has been described as "appropriateness of the test development process", and this refers to "all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material" (Sireci & Faulkner-Bond, 2014, p.101, cited in Hoeve,2022, p.3)**.** It is said to be at the heart of the validation process called "alignment", which "involves evaluating the correspondence between student learning standards and test content" (The Standards, p.15, quoted in Hoeve, 2022, p.3). Scholars are thus in agreement that content validity is the core of validity. There are also in agreement that content validity concerns the extent to which a test's content reflects the stipulations of the syllabus or covers the specific knowledge area and language skills requirement to be tested. However, there is no consensus on how to establish content validity.

For Lado (1961), the chief proponent of Contrastive Analysis (CA), establishing the validity of the content of a language test would take the form of a linguistic analysis to determine whether or not it contains the learning problems. The assumption is that testing control of the learning problems is testing the language. Accordingly, a test lacks validity if its linguistic content fails to reflect testees' learning

problems. Criticisms against CA are well catalogued in the literature and need not be rehashed here. However, for the sake of emphasis, it is necessary to restate the undesirability of basing ESL tests solely on previously identified learning problems. These are largely mother tongue dependent variables and are difficult to sample in a multi-lingual speech community like Nigeria, where what constitutes learning problems to speakers of Language A would be a walk-over test for Language B-speaking testees. Probably because of his quick realisation that all learning problems are not necessarily mother tongue-derived, Grieve (1964) suggested that the "kind of language needed to cope with practical situations in everyday life" (p.21) should be tested, in addition. This was unfortunately left undefined or clarified.

Harris (1969) held a contrary and more comprehensive view on how to establish content validity for language tests, For him, the validity of a language test's content should be "based upon a sound analysis of the skill or skills we wish to measure", with "sufficient evidence that test scores correlate fairly highly with actual ability in the skills areas being tested" (p.19). Though Harris failed to specify how to determine the skills appropriate for the level at which we are testing (and no one after him appears to have done that either), this position of his presupposes two vital elements of language testing. These are (a) having clearly defined goals to guide the determination of test content and (b) developing an appropriate teaching or examination syllabus against which the linguistic content of tests can be evaluated. The aspect of testing goals requires critical scrutiny in view of the gradual but steady shift "from assessment of learning to assessment for learning" (Osidak et al., 2021, p.2). This is a dangerous trend in modern day language testing that should be checked. The foregoing underscores the significance of content in language test evaluation, despite the divergent opinions on what constitutes a language test's content and how it can be appropriately sampled.

Scoring validity

The term scoring validity was introduced by Weir (2005) as a superordinate term to reflect the fact that, "following Alderson (1991a), reliability is now increasingly seen as a type of validity evidence" (p.23). It is understood as reliability with a new identity, and it concerns the extent to which test results are stable over time, consistent in terms of content sampling, and free from bias. In other words, scoring validity accounts for "the extent to which examination marks are free from errors of measurement and therefore the extent to which they can be dependable for making decisions about the candidates" (p.23)**.** This notion of scoring validity will be invoked in respect of ESL Objective Tests' validity and how best to determine it.

Methods of investigating validity

Three methods of investigating validity are discernible from the literature. One is essentially statistical, and it involves (a) analysis of internal test structure (to report statistical relationship among parts of the assessment), (b) exploring relations to other variables/constructs (to report statistical relationship among reported scores on the total test and subtest), and (c) test speededness (evaluating differential effects of speed tests, say on learners with low reading fluency). These procedures are premised on the realisation that a critical aspect of validity is to determine whether "a test is measuring the same construct across the different sub-populations" (ELT, 2009, p. 27)**.** Another method is the corpus-based approach, adopted by Pan and Quian (2017) to investigate the content validity of the grammar section of a language test in Shanghai, China. They reported a fairly average content coverage, and proposed a procedure for conducting corpus-based content validation of high-stakes grammar tests such as WAEC's SSC English Language Objective Tests.

What may be regarded as a third method is actually a reaction against a wholly statistical approach to construct validation. Weir (2005) noted that theory-based validity comprises both *a priori* and *a posteriori* components and that very little attention was paid to the non-statistical aspects because construct validity was viewed "principally as a posteriori statistical validation" in much of American tests (p.18). Warning that "statistical data themselves do no generate conceptual labels", Weir expressed the need for validation at the a priori test development stage in order to make the a posteriori validation more meaningful. Furthermore, he argued that it is important to describe the construct to be measured as it is to investigate the adequacy of the testing procedure. Considerations for a priori test validation would give linguistic input its rightful place in the validation process, with greater relevance for ESL Objective Tests.

Summary

The foregoing shows that language test validation for most scholars is scores-oriented, with very little consideration given to linguistic input. It is therefore highly limited. There are also no criteria outlined specifically for evaluating objective tests, despite their popularity and peculiarity. The general criteria seen in this review cannot adequately serve the purpose of evaluating ESL Objective Tests because of flaws which the next section will attempt to redress.

Validating ESL Objectives Tests: A case for greater linguistics input

The paradox of the objective technique (which concurrently qualifies it as the most popular testing format and the most unsuitable for testing actual language use ability) necessitates setting up special criteria for its evaluation. This is the concern of this fourth section: It articulates three criteria for the validation of ESL Objective Tests based on the underlying premise that the tests are fundamentally language tests and should technically be evaluated with necessary and sufficient input from linguistics (the

discipline that studies language scientifically). This task also acknowledges the need to clearly define the goals of testing ESL or any of its more specific aspects as starting point. These must reflect the skills areas, the appropriate language items that go with them, and the peculiarities of the ESL situation. As a matter of fact, an ESL test is valid only if its contents can lead to the realisation of ESL testing goals.

Before proceeding further, it is necessary to propose that validity be used as an all-embracing, superordinate term for language test evaluation. This proposal is based on the argument that, whereas reliability and practicality are generally regarded as necessary conditions for validity, validity is neither a condition for reliability nor practicality (Adejare, 1987; 1991). Bringing all three criteria under the terminological umbrella *validity* has the advantage of providing for greater linguistic input into the process of validating language tests generally. And, it needs to be emphatically stated that greater linguistic input does not necessarily undermine other sources and types of validity evidence exemplified by scoring validity. On the contrary, it complements them. The proposal for the adoption of validity as a cover term for language test evaluation is not without theoretical justification, as already seen. Further justification is provided as follows.

Bachman and Palmer (1996) noted that reliability and construct validity are "sometimes referred to as essential *measurement* qualities" because they "provide major justification for using test scores" (…) and for "making inferences and decisions". They further noted that reliability is a necessary, but not a sufficient, condition for construct validity and hence usefulness because, while objective tests are very consistent and reliable, they cannot justifiably serve as basis for writing course placement (p.23). This can be interpreted to mean that validity is not a condition for reliability; it is not a condition for practicality either. In contrast, poor testing considerations can impact negatively on validity. Weir (2005) similarly argued that the traditional polarisation of reliability and validity is unhelpful and that reliability is better regarded as a form of validity. He had earlier explained that validity is multifaceted and that no single validity is superior to the other, as "deficit in any one raises questions as to the well-foundedness of any interpretation of the scores" (p.13). This was the point at which Weir disagreed with Messick (1995,742) who, he pointed out, rightly held a comprehensive view of validity and also recognised the complementarity of its different sources, but who upheld construct validity as a superordinate category for validities. Weir jettisoned the idea of construct validity (which specifically refers to the theoretical construct--individual cognitive ability) as a superordinate term for all the validities, but preferred instead to reinstate the term validity as the superordinate category. The use of the verb *reinstate* suggests that validity had probably been long recognised and used as an umbrella term for test evaluation (Attempts at verifying this claim have been futile). So, validity lies in the heart of language testing generally, and it is "one of the most important attributes of assessment" ETS (2009).

Having made these remarks, the linguistics-informed criteria can now be identified. They are linguistic content, testing technique, and testees' performance. The rest of this section explicates the three broad criteria in that order.

Linguistic content

Since an ESL objective test is intrinsically a language test, the first criterion for its evaluation is naturally linguistic content. Linguistic content generally covers the grammatical, lexical, and phonological levels of linguistic description and includes the four language skills of listening, speaking, reading, and writing. Consideration for linguistic content specifically entails examining the adequacy of coverage, linguistic adequacy of individual test items, and the effectiveness of distracters.

Adequacy of coverage

The first and most critical question to ask at the initial stage of the evaluation process is this: How adequate is the test coverage with respect to (a) the teaching or examination syllabus and (b) the various units of grammar, lexis, and phonology and the relevant skills areas? Where a test is based on a syllabus, the task is relatively easy. All that needs to be ascertained is whether or not the content and specifications of the syllabus are adequately reflected in the test. On the other hand, where a test is not overtly based on a recognisable or prescribed syllabus, a good validation job can still be done by assigning individual test items to the various descriptive units of grammar, lexis, or phonology. This will not only allow a clear picture of the test's coverage to emerge, but will also create room for comparative weighting in respect of linguistic levels such as grammar and lexis, the two most commonly co-tested levels in standardised ESL Objective Tests. It should be noted that this last procedure may not be emphasised for tests based on a syllabus because of the assumption that a good syllabus would have met the adequacy of coverage condition before it was approved for use, except where the syllabus itself is being evaluated. Two more factors must be brought in at this stage. These are ESL testing goals and linguistic models. For all tests, it needs to be ascertained that their contents are capable of leading to the realisation of the defined goals of ESL testing, and this is irrespective of whether or not there is a syllabus.
The second equally important factor is linguistic model. Linguistic models so greatly influence language tests that Strevens (1965) and Harris (1969) suggested that they be made a condition for determining the validity of language tests. An ESL test must not only be based on an appropriate linguistic model for validity, it must also be seen to reflect the model consistently throughout. The need for appropriateness in the choice, and consistency in the use, of a linguistic model assumes a more critical

dimension in ESL Objective Tests because of the numerous criticisms against the objective technique. Although it could be somewhat difficult to determine precisely which linguistic model informs a particular objective test by merely examining its content, a good result can nevertheless be obtained if features such as contextualisation and descriptive nonces are considered. The extent of coverage and relative weighting are also indices of linguistic models. There are strong indications that Systematic Grammar is beginning to exert more influence on language testing today. This is a positive development in light of its insistence that language should be studied in relation to its users and uses, a theoretical position that tacitly recognises the ESL variety and that makes the systemic grammatical model the most suitable for ESL testing. It is equally not difficult to notice that vestigial traces of Traditional Grammar and Bloomfieldian structuralism menacingly pervade the tests. This is evidence of test writers' unwillingness, reluctance, or inability to completely shake off centuries-old influence of Traditional Grammar on language testing.

Linguistic adequacy of test items

Next to adequacy of coverage is linguistic adequacy of individual test items. Here the test validator is concerned basically about five things. First, there should be only one clearly defined goal for every test item (This is different from defining the purpose for each test, which the testing entity ETS suggested). Second, each test item should be assignable to only one identifiable linguistic level, unit or sub-unit. Third, appropriate and adequate contexts must be provided for each test item. Fourth, every test item must be absolutely free of errors (grammatical, phonological, lexical, or graphological). Finally, every test item must be free of factors extraneous to linguistics.

Effectiveness of distracters

Closely linked to adequacy of test items is effectiveness of distracters. This is a condition for the overall effectiveness of the individual test item in a multiple-choice objective test, and it is determined statistically by finding out the number of times a distracter is chosen by testees (Harris, 1969). Whether or not a distracter is chosen contributes to determining how effective or ineffective it is. For instance, a distracter not chosen at all, or chosen only by a few, is said to be ineffective. On the other hand, a distracter that attracts a fairly large number of respondents is adjudged effective because it is able to distract testees from the answer. But there are obvious contradictions inherent in the use of Harris' method of determining the effectiveness of distracters. For example, an item may be easy, thereby rendering the distracters ineffective. Alternatively, the distracters may be effective when the item is difficult. Indeed, they may be effective or ineffective when they are structurally defective or when they manifest visual clues. Given the absence of clear-cut criteria, it is hereby advocated that linguistic criterion be used as an additional measure towards determining the effectiveness of distracters. This should have priority over any other set of criteria. Thus, each distracter can be examined to determine its linguistic relatedness to the item tested (syntactically, morphologically, phonologically, lexically, or semantically) and to ascertain whether it meets the grammaticality and acceptability conditions, as deviant structures or errors reduce rather than enhance the validity of test items.

Testing technique

The second major linguistics-informed criterion for ESL Objective Tests evaluation is testing technique. The question that one intuitively anticipates in this regard is: Why evaluate technique when the test's name implies a testing technique? Notwithstanding this observation, it is considered absolutely necessary to evaluate the technique used, for a language test cannot be fully certified as valid if the testing technique employed is faulty. Basically, the sub-types of the multiple-choice objective technique are sentence completion and sentence interpretation. There is also the cloze procedure, used.in its objectified form to test lexis (register). An application of the technique criterion demands that satisfactory answers be provided to at least the following questions:

a) Is the technique adopted suitable for the type of skills area or language structure tested?
b) Are the testees handicapped or disadvantaged in any way by the use of the said technique?
c) Will testees' performance in the skills area or language aspects tested be considerably hindered or enhanced by the specific mode of testing employed?
d) Does the technique enable the tester to cover as many grounds as are necessary without repetition, negligence, under-representation, or over-concentration?
e) Does the test format enable the tester to focus clearly on the defined goal without the test being unnecessarily difficult, over-simplified, or even ambiguous?

In dealing with technique as a criterion, the evaluator should be careful not to introduce unrelated instruments such as the non-linguistic content of a cloze passage.

Testees' performance

ESL Objectives Tests can also be evaluated using testees' actual performance. as basis. Testees performance is used to determine the level of difficulty through item analysis (Ingram, 1968), based on the rule that language test items must be of moderate level of difficulty. Raw scores obtained can also be analysed to determine consistency and thereby validity. Perhaps of greatest significance for evaluation is the correlation of testees' performance in ESL Objective Tests with their actual ability to use the language, precisely the skills areas or linguistic structures covered by the test. This can easily be done by assigning a written or oral composition topic, where testees are required to employ to a great extent some of the linguistic structures tested (Adejare, 1987). The correlation could be high, medial or low and each outcome has implications for the overall validity of the tests, especially from the perspective of the largely empirically unsubstantiated claim that the objective test is not an adequate test for actual language use ability.

Conclusion

In this paper, I argued for greater linguistic input into the evaluation of ESL Objective Tests to enhance the

tests' validity. I particularly argued that, since reliability and practicality are necessary conditions for validity whereas the corollary is untrue, validity should be made an all-embracing, superordinate term for language test evaluation. Following the adoption of validity as a cover term, I extensively reviewed the concept and demonstrated that its interpretation and use in the literature on language testing is invariably statistics-based. Consequently, I articulated three broad linguistics-informed criteria for establishing the validity of ESL Objective Tests that do not necessarily jettison the existing statistics-based criteria. These comprise linguistic content, testing technique, and testees' performance. With particular reference to linguistic content, I argued that its capacity to realise ESL testing goals rather than identified learning problems should form the basis for its determination and evaluation and that this should be prioritised in order to mitigate the perceived inadequacies of ESL Objective Tests. I also advocated the return of Spoken English to ensure consequential validity and improve the tests' quality. A combination of statistical criteria with linguistic criteria would produce the type of results desired for the validation of ESL Objective Tests and indeed language tests in general. Language test evaluation deserves nothing less.

References

Adejare, R. A. (1987). English language testing: An evaluation of the West African school certificate English language objective test. Unpublished M.A. ESL Thesis, University of Ife.

Adejare, R. A. (1991). Evaluative criteria for ESL objective tests. Paper presented at the 22nd Annual Conference of the Nigeria English Studies Association (Silver Jubilee Conference), University of Ibadan, Ibadan, 16-21 September 1991.

Adewumi, A.A. (1996). Towards functional examining in English language. *The Nigerian Language Teacher.* 6 (2), 50-53.

Aryadoust, V., Zakaria, A., Lim, M.H., & Chen, C. (2020). An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Frontiers in Psychology*. 11:1941.162-190 https://doi:10.3389/fpsyg. 2020.01941.

Bachman, L. F. & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bustamante, A. (2022). Language assessment literacy of middle school English teachers in Mexico. *Languages 7* (32), 1-21. https://doi.org/10. 3390/languages 7010032, 11 www.mdpi.com/ journal/languages.

Educational Testing Services (ETS) (2009). *The guidelines for the assessment of English language learners*, pp. 1-42. Authored by M. J. Pitoniak, J. W.Young, M. Martinielo, T. C. King, A. Buteux, & M. Ginsburg. (A Document) www.ets.org

Grieve, D.W. (1964). *English language examining*. African University Press.

Harris, D.P. (1969). *Testing English as a second language.* McGraw-Hill.

Hoeve, K.B. (2022). A validity framework for accountability: Educational measurement and language testing. *Language Testing in Asia.* 12 (3), 1-14, Springer Open, https://doi.org./10. 1186/s40468-0400153-2,

Ingram, E. (1968). Item analysis. In A. Davies (Ed.). *Language testing symposium: A psychological approach.* Oxford University Press.

Ivanova, V. & Terzieva, T. (2015). Criteria for the construction of tests for language assessment and evaluation. Conference Paper. 58-65. https://www. researchgate.net/publication/ 329000142.

Le, Vi-Nhuan. & Klein, S. P. (2002). Technical criteria for evaluating tests. In L. Hamilton, B.M.Stecher, & S.P. Klein (Eds), *Making sense of test-based accountability in education*. 51-79. Rand Corporation. https://www.jstor.org/ stable/107249/ mr1554edu.10.

Lado, R. (1961). *Language testing.* Longman.

Osidak, V., Drahinda, O. & Kvasova, O. (2021). Training the trainers in language assessment via mentoring: Building expertise to promote language assessment literacy of Ukranian university teachers. *Languages,* 6 (194), 1-21. https://doi.org/10.3390/languages. 6040194.

Pan, M. & Quian, D. O. (2017). Embedding corpora into the content validation of the grammar test of the national matriculation English test. (NMET) in China. *Language Assessment Quarterly.*14 (2),120-127. https://doi.org/10. 1080/15434303.201713 03703.

Saglam, G., Lidice, A. & Tsagari, D. (2022). Evaluating perceptions towards consequential Validity of integrated language proficiency assessment. *Languages*. 7(65), 1-18. https://doi.org/ 10.3390/languages7010065

Sheeba, M.R. (2020). Testing and evaluation in English. *India Review of World Literature in English*. 16 (11), 1-7. https://worldlitonline. net/.

Strevens, P.D. (1965). *Papers in language and language teaching.* Oxford University Press. *The Oxford Advanced Learners Dictionary of Current English* (OED). Oxford University Press.

Weir, C. J. (2005). *Language Testing and Validation. An evidence-based approach.* Research and Practice in Applied Linguistics. Palgrave Macmillan.

Note

This paper was presented at the 22nd Annual Conference of the Nigeria English Studies Association (Silver Jubilee Conference) held at the University of Ibadan, Nigeria between 16-21 September 1991 as "Evaluative Criteria of ESL Objective Tests". It was subsequently accepted for publication in *Journal of the Nigeria English Studies Association,* Vol. 11, Nos. 1& 2, 1994 but that publication never saw the light of day for some reasons. It was extensively revised for this volume because of the strong conviction that the issues raised then are as relevant and topical today as they were some three decades ago.